DOCUMENT RESUME

ED 062 647 CG 007 168

AUTHOR	Wexler, Kenneth N.; And Others
TITLE	Structure and Process in Cognition. Workshop Report.
INSTITUTION	California Univ., Irvine.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-8-0293
PUB DATE	15 Aug 70
NOTE	383p.; CORBE Research Workshop on August 16-21, 1970 in Huntington Beach, California
EDRS PRICE	MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS	*Cognitive Development; *Cognitive Processes; *Language Research; *Psycholinguistics; *Psychological Studies; Structural Linguistics; Thought Processes

ABSTRACT

Nine papers were presented at a workshop concerned with structure and process in cognition. The reports given at this conference represent detailed applications of these concepts to limited areas of experience. This work is unique in that both structure and process appear together, rather than being studied as independent concepts. All the papers may be interpreted as talking about processes that apply to structure. This represents the beginning of a new kind of work in cognitive psychology. These papers are directly concerned with semantics. This suggests that semantics is providing a setting for further work in cognition. The workshop did not produce any specific conclusions of findings. The interrelations between the structures and processes studied in each of the papers have been worked out. Each paper analyzes a limited range of phenomena. There is not one view which at the moment can cover most of the phenomena. Rather, these papers may be the basis for the future development of an adequate general theory of cognition. [Not available in hard copy due to marginal legibility of original document.] (BW/Author)

ED 062 647

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

COBRE Research Workshop
on
COGNITIVE ORGANIZATION AND PSYCHOLOGICAL PROCESSES

August 16-21, 1970
Huntington Beach, California

ED 062 647

CG 007 168

COBRE Research Workshop

on

COGNITIVE ORGANIZATION AND PSYCHOLOGICAL PROCESSES
August 16-21, 1970
Huntington Beach, California

WORKSHOP REPORT "STRUCTURE AND PROCESS IN COGNITION: Report on the COBRE Research Workshop on 'COGNITIVE ORGANIZATION AND PSYCHOLOGICAL PROCESSES'" by Kenneth N. Wexler

1. "The Use of The Balanced Block Design for the Triads Test of Judged Similarity" by Michael L. Burton and Sara B. Nerlove

2. "How We Understand Negation" by Herbert H. Clark

3. "The Psycholinguistic Structure of Knowledge: 1. Some Examples" by Edward J. Crothers

4. "Cognitive Structures and Judgement" by Roy G. D'Andrade

5. "A Theory of Marking Rules" by William H. Geoghegan

6. "Utilization of Cognitive Structures in Problem Solving and Reasoning" by James G. Greeno

7. "Semantic Structure and Clustering in Free Recall" by W. Kintsch, R. F. Miller, and R. M. Hogan

8. "Toward a Theory of Analogical Reasoning" by David E. Rumelhart and Adele A. Abrahamson

9. "Semantic Structure: Psychological Evidence for Hierarchical Features" by Kenneth N. Wexler

CG 007 168

Structure and Process in Cognition

Report on the COBRE Research Workshop on COGNITIVE ORGANIZATION AND PSYCHOLOGICAL PROCESSES

(August 15-22, 1970)

Kenneth N. Wexler
University of California, Irvine

The notions of structure and process underlie current work in the cognitive sciences. In particular, the reports given at this conference represent detailed application of these concepts to limited areas of our experience. What is unique about such work, I think, is the extent to which considerations of both structure and process appear together. Up to recently, psychologists, anthropologists and linguists have tended to study one or the other. One justification for excluding one of these notions is that it is not important, or not relevant, or not 'real'. That is, the notion is not needed. A second justification is a division of labor argument. Given limited time, it is possible to study either structure or process, but not both. Nevertheless, both are relevant.

I think that it is becoming clear that neither of these arguments is adequate. Not only are both structure and process important to study, but they cannot be studied independently without losing a great deal of insight. While I say that this situation is unique or new, I mean this in only a very local sense, that is, given the recent history of the cognitive sciences.

In this connection it is interesting to consider what Alfred North Whitehead had to say on the subject. In Process and Reality (1929) he notes the saying of Heraclitus that 'all things flow' and writes "that 'all things flow' is the first vague generalization which the unsystematized, barely analyzed intuition of men has produced... Without doubt, if we are to go back to that ultimate, integral experience, unwarped by the sophistications of theory, that experience whose elucidation is the final aim of philosophy, the flux of things is one ultimate generalization around which we must weave our philosophical system... But there is a rival notion, antithetical to the former. I cannot at the moment recall one immortal phrase which expresses it with the same completeness as the alternative notion has been rendered by Heraclitus. This other notion dwells on permanences of things..."

Whitehead considers how the two notions are expressed together in the lines

Abide with me;
Fast falls the eventide.

He writes "Here the first line expresses the permanences, 'abide, 'me' and the 'Being' addressed; and the second line sets these permanences amid the inescapable flux. Here at length we find formulated the complete problem of metaphysics. Those philosophers who start with the first line have given us the metaphysics of 'substance'; and those who start with the second line have developed the metaphysics of 'flux.' But, in truth, the two lines cannot be torn apart in this way; and we find that a wavering balance between the two is a characteristic of the greater number of philosophers."

A fundamental point here is that both structure and process are to be studied and that they are intimately related. Actually Whitehead comes down on the side of process as being more fundamental. I don't want to imply agreement with Whitehead's particular philosophy. The above quote is more or less a historical description.

In fact, as our theories become more and more precise, it becomes clearer how they can fit into the mode of this kind of historical description. Let me quote just one more sentence from Whitehead. He writes, "On the whole, the history of philosophy supports Bergson's charge that human intellect 'spatializes the universe'; that is to say, that it tends to ignore the fluency, and to analyze the world in terms of static categories." One immediately thinks of multi-dimensional scaling. If we provide a Euclidean representation as a model of the structure of a set of objects to which humans respond, we have "ignored the fluency", that is, we haven't discussed the processes that people go through in responding to these objects. Of course, the opposite occurs, that is, theories have been advanced which concentrate on process to the exclusion of structure. I would argue that traditional learning theory may be described in this way.

There is another distinction which correlates with the one already made. This involves the difference between 'natural' and 'artificial' structures. Anthropologists and linguists have concentrated on discovering natural cognitive structures and have tended to ignore process. Psychologists have concentrated on process, for example, memory and learning, while dealing with artificial, laboratory-induced structure. Doubtless a reason for this is one of relative simplicity. If one is studying process, the problem is difficult enough without becoming involved with complicated structures. Of course, one may then pay the price of not being able to say much about structures that people actually use. A theme running through the papers at the Workshop is an attempt to study both natural structures and natural processes.

In this report I want to discuss each of the 9 papers in order to see how they exemplify this general notion of structure and process. In order to do this I will have to first give a short summary of the paper. I will occasionally report comments made at the Workshop but not very often. Only those comments that seem to speak to general issues will be reported. Discussions of technical points, for example the relating of response time results to those obtained in other experiments, will be omitted. The papers will be discussed in the order in which they were presented at the Workshop.

Rumelhart and Abrahamson

The paper by Rumelhart and Abrahamson, "Toward a Theory of Analogical Reasoning", provides a nice example of the role of the notions of structure and process in cognitive theories. Subjects are asked to perform analogies on animal terms, for example, "beaver" is to "sheep" as "dog" is to ______, with the subject required to fill in the blank term from the 4 alternatives given, in this case, "donkey", "camel", "elephant", and "chimpanzee". The authors consider analogy as a kind of reasoning and remark that "...the theoretical problem in understanding any particular reasoning process becomes clear. We must (1) specify the form of the memory structure and then (2) determine the algorithm which is applied in the case of the reasoning process in question".

It is assumed that the memory structure for the animal terms is a Euclidean space, with each term represented as a point in the space. The specific space is 3-dimensional and is taken from Henly (1969), who derived it by applying a multi-dimensional scaling program to judgments of similarity on the animal terms. Between each pair of points in this space there can be calculated a directed (vector) distance. It is assumed that for any analogy of the form A:B::C:? there is an ideal analogy point I such that the vector distance from C to I is the same as the vector distance from A to B. It is also assumed that the closer (in absolute distance) an alterntive is to I, the more likely it will be chosen as the best analogy solution.

Three experiments were performed to test the theory. In the first two, subjects did analogies on animal terms. The basic assumptions of the theory were confirmed, that is, the further from the ideal point an alternative was, the less likely it was chosen as the solution. Subjects were also asked to rank the four alternatives. With three added assumptions, including Luce's choice rule, predictions were also made of the probabilities of ranking of each alternative. These predictions were also considered confirmed.

In the third experiment, an attempt was made to teach concepts by means of analogies. Thus 3 sense syllable points were chosen in the animal space and labeled BOF, DAX, and ZUK. Subjects were then given analogy problems of the form, for BOF, say, of $A:B::BOF:(X_1, \ldots, X_4)$. The subjects guessed an answer and then were informed of the best answer and the rankings of all the alternatives. Following this learning, subjects rated (on a scale of 1 to 10) the similarity of each of the 3 new "animals" to the original 30 animals and to each other.

After the fifth trial of learning subjects responded to the analogy problems in a way predicted by the theory. That is, their probabilities of ranking of the alternatives were quite similar to what the theory would predict for animals at the points assigned the new "animals". The similarity ratings of the artificial animal terms also behaved as they should given their distances from the other animals. Thus this experiment provides an illustration of concept fromation via analogical reasoning.

Discussion of Rumehart and Abrahamson's paper centered around the structure and process parts of the theory. It was claimed that important semantic relations which people use in analogies are not represented in the theory. D'Andrade discussed the theory in relation to the work of Evans (1968) and claimed that only part of the analogical reasoning process was represented in Rumelhart and Abrahamson's theory. Important preliminary processes, such as determining features and relations between features were not considered in the theory.

If we take a Euclidean space as a model of the subjects' representation of the set of animal terms, then the subjects' solutions of analogy problems can be predicted. But it seems to me that this result does not provide evidence for the proposition that an adequate model of the animal terms is a Euclidean space. Rather it implies that the processes of solving anaolgies and of making similarity judgments are closely related. Thus if we have information on similarity judgments (the Euclidean space derived from those judgments) then we can make some predictions about analogies. The wider the variety of tasks that can be understood by applying various process models to the Euclidean model of structure, then the more likely we are to believe that the model is a representation of subjects' knowledge.

Kintsch, Miller and Hogan

The paper by Kintsch, Miller and Hogan is a study of the relations between the cluster structure obtained in free recall experiments and that obtained in tasks which are directly dependent upon semantic organization. Kintsch et al. call the component of memory that represents the meanings of words the 'lexical component of memory' and propose the hypothesis that this component 'provides the basis for clustering in free recall'. The alternative hypothesis is that clustering in free recall is not related to the general memory structure.

Three tasks were used to measure the similarity of words in long-term memory (that is, the lexical component). The first method was to ask subjects to sort words into categories on the basis of their similarity of meaning. The more often that two words occurred in the same pile, the more similar they were. (Two measures were obtained from this task, one from before and one from after "correction" of the sorting of the subject). The second method was to ask subjects to identify a word (selected by the experimenter and unknown to the subject) by asking questions about it. The number of questions answered "yes" that are identical for two words provides a measure of similarity between the words. The third measure of semantic similarity was obtained by measuring the response overlap to two different words in a restricted association task, that is, the number of associates the two words had in common.

These tasks were all done on the same set of 40 nouns. A free recall experiment was carried out on the same words. The measure of similarity between two words was obtained by counting how often a word was immediately followed by the other word in the responses of the subjects. This measure was obtained for both trial 1 and trial 5.

In order to analyze the similarity of these measures, clusterings were produced from each of the 6 proximity matrices (2 sorting measures, identification task, association and trials 1 and 5 of free recall task). A clustering is simply a partition of the set of words (it is not a hierarchy). A measure of the similarity of two clusterings was developed and calculated for each of the $\binom{6}{2} = 15$ pairs of clusterings.

In all 15 cases the value of this similarity measure between clusterings was quite high. Looking at the clusterings produces the same result. When there is a difference between 2 clusterings it tends to be the case that one partition is simply finer than the other, that is, one of the clusterings may break a subset of words into, say, 2 clusters, while the other treats them as one cluster. This result held for all 6 clusterings, including the 2 obtained from the free recall data. The conclusion is that there is much similarity between lexical organization and clustering in free recall. (A second experiment, similar to the first but using more highly structured sets of words, produced essentially the same results.)

What are 'structure' and 'process' in this study? As the authors admit, they are not well specified. The 'structure' is the structure of the lexical component of memory but that is not given. As the authors write, 'Although we do not know and we do not have a model for it, we can compare the output order in free recall with performance in various other tasks in which lexical structure may reasonably be expected to be a crucial factor... . Although the limitations of this empirical approach are obvious, it can provide evidence as to the general adequacy of our working hypothesis. If the structure of the output order in free recall does not correlate well with that obtained in the other tasks mentioned, the usefulness of the hypothesis that the output order in free recall reflects, *inter alia*, lexical structure would be questionable'.

The clusters are not a structure in the sense in which we have used the word 'structure'. Rather they are the results of task-specific processes working on some unknown structure. The processes are also unspecified. Miller (1969) has discussed some models of structure and process for the sorting task, and there are many models of free recall in the literature which, however, either do not take the structure of the materials into account, or assume that the structure is given by a numerical measure of association between 2 words. The problem is to develop models of lexical structure which allow process models to be defined on them and predictions made for both semantic similarity and free recall experiments.

Wexler

Wexler's paper is an experimental study of semantic structure. It is an attempt to develop a psychological theory of the representation of semantic features. As an experimental example, a set of 9 'have' or 'transfer' verbs was analyzed, including, for example, 'have', 'get' and 'give'. In order to determine the organization of the meaning of this set of verbs, a triads test was performed.

In this test, the subject is given 3 words and asked to select the one which is 'most different in meaning from the other two'.

It turns out that the results of this experiment can be predicted fairly well by the following model. The structural model is a tree, with the non-terminal nodes labeled by semantic 'features' and the terminal nodes representing the 9 words. The first feature is whether or not the subject has the object after the action, and the second feature is whether the subject has the object before the action. The process model, which determines the choice in the triads test, applies to the tree. It looks for a difference in the 3 words on the first feature. If there is a difference a decision is made then. That is, the word which has a different (from the other 2) value on that feature is selected as most different. If the value of the first feature is the same for all 3 words, however, then the model goes on to the second feature in the tree. The process is repeated until a difference is found on some feature.

In addition to predicting the triad choices, the model can be made to predict relative response times for triads. If we assume that each search of a feature takes some amount of time, then triads for which the process model can come to a decision after the first feature is searched should take less time than triads for which the model must search 2 features, and so on. For this particular structural model (tree), 3 sets of triads can be distinguished and ranked according to predicted response times. An experiment measuring response times was performed, and indeed, the predicted rank order amoung the average values for the 3 sets was observed. In fact, the response times appeared linear with respect to the depth of the feature at which a decision could be made, a result which would be predicted if we added the assumption that each feature took the same amount of time to process.

If one now carries out the process model on a somewhat more minute level, that is, attempts to spell out a few of the assumed processes in more detail, it turns out that the tree model is not as likely as a lexicon model, that is, a model in which each word has stored next to it a list of semantic features. The evidence for this is also in the response time data. Also, the data tend to rule in favor of a top-down rather than bottom-up model of processing the features.

One other experiment was performed, a triads test in which the items were sentences with constant frames containing the verbs rather than just the verbs. This experiment was done in order to compare the results of the analysis with the structure obtained by Bendix (1966) for the same set of words in frames, using a linguistic analysis. His structure did not do nearly so well in predicting the data as a hierarchical structure essentially the same (but allowing for the frames) as the one used to predict the results for words in isolation.

The models described in this work must meet the same objections that apply to most of the other models we are considering. Namely, they have been developed to fit a particular situation (the triads test) and have not yet been tested in other situations. In order to be considered more generally valid, process models have to be developed which apply to the same structure and predict what happens in the new experiments. One can think, for example, of obvious models which apply to trees to explain analogy experiments of the Rumelhart and Abrahamson type. But it remains to be shown whether they can do as well as Euclidean models.

A number of comments were made concerning the fact that the implication relationships in the set of verbs were not explicated by the model. For example, 'take' implies 'get'. In order to account for these relationships, one could construct a tree with words allowed at all the nodes, not just the terminal nodes. Then if A dominates B in the tree, this is to be interpreted as B implies A.

A question was raised as to why the tree had to take its particular form. Why, for example, couldn't the first feature be whether the subject had the object before the action and the second feature relate to what the state of affairs is after the action? There are 2 ways of answering this question. The first is that, given the process model, the data is predicted quite a bit better by constructing the tree as it has been constructed. The second answer is that, given this is the case, why is it so? I can think of 2 tentative answers to this. The first is that ours is a materialistic culture, and we are interested mostly in what the state of affairs is with respect to possession now, at the completion of the action. A second answer, which I prefer, is that features are stored in the order in which they are acquired and that the after feature is acquired first because it is most available to the child at the moment that he is trying to learn the meaning of the word. For example, I say 'Daddy gave Mommy the book', and the child sees that now Mommy has the book. In fact, a developmental experiment that I have done provides evidence that the features in the tree are learned in order, starting from the top. Lower features are learned later because they do not involve physical action. In developing a semantically-based theory of language acquisition, Peter Culicover and I have had to conclude that physical action is highly salient to the child.

D'Andrade

D'Andrade's study was an attempt to show that a method generally used in psychology to study human personality is invalid because the method relies on human judgment from memory, which is 'subject to distortion or bias in the direction of pre-existing cognitive structures.' The general method is to have humans judge subjects on some traits and then to compute correlations for all pairs of traits on these scores. The pattern of correlations are then analyzed to determine which traits go together, that is, to determine, for example, clusters of traits.

The ingenious method used in D'Andrade's 2 analyses was to show that the memory-based judgments of human observers showed a pattern of correlations among the traits that was more similar to the pattern of 'semantic similarity' judgments of the names of the traits than to the pattern determined by judgments made by human observers which did not depend on memory, that is, judgments of behavior recorded as the behavior took place. The first analysis was based on data collected by Borgatta, Cottrell and Mann (1959). In that study, subjects in a class were brought together for 'discussion'. After weeks they ranked each member of their group on a number of traits. During some of the sessions a trained observer recorded the behavior of the subjects according to the Bales category scheme. Six of the categories were similar to one of the traits used in the human rankings, and thus these 6 categories were analyzed. An example of a category is 'shows solidarity and friendliness'. Another example is 'makes the most suggestions'.

D'Andrade performed a semantic similarity test on these (slightly modified) categories, asking subjects to judge the similarity of a pair of categories on a scale from +3 (very similar) to -3 (very dissimilar). For example, subjects were asked to judge the similarity of 'shows solidarity' and 'suggests, gives direction'.

There are now 3 correlation matrices, that is, tables ofcorrelations between traits. The first is for behavior rates, as determined by the subjects, and the third is for semantic similarity ratings, as determined by experiment. Now, in order to determine the similarity of these matrices, correlations betwen the entries were computed for each pair of matrices. The important finding is that the correlation between the rank judgments and the semantic similarity ratings is that the correlation between the rank judgments and the behavior rates. In other words, which traits 'went together' in the memory-based human judgments did not reflect so much which traits actually went together in behavior (or, at least, direct observation of behavior), but reflected the 'pre-existing cognitive structure'.

Other analyses were done of the same data, and a second set of data (Mann, 1959) was also analyzed. The basic results were the same. It should be noted that in this second study the observer who made the immediate recording of behavior also judged the subjects from memory, after the session. The intriguing result is that the pattern of correlations of traits of the observer's memory-based judgments was more similar to the pattern for semantic similarity ratings than to the pattern for his own immediate recording of behavior. This result provides evidence against the possibility that the results in the first study were found because the Bales categories do not mean the same thing to the observer as to the subjects who make the judgments.

What can we say about structure and process in D'Andrade's study? These aren't spelled out in the paper but we can speculate about them. In the first place, there is forgetting. That is subjects don't remember how people behaved. This is demonstrated, for example, by the fact that in the second study the median correlation (over categories) between the immediately recorded behavior and the memory-based judgments for a category was less than .36. That there really was forgetting and not simply failure to perceive the behavior, is indicated by the fact that this result is not much improved even when we consider only the memory based judgments of the same observer who made the immediately recorded behavior judgments. There are many models of forgetting, of course, and there really is no way of suggesting, given the current data, which is most appropriate.

But as D'Andrade points out, not only is there "memory drift when people make ratings or rankings of other people's behavior, but . . . this drift is systematic, nonrandom, moving in the direction of the rater's conception of 'what is like what'". (One thinks, of course, of the famous work of Bartlett (1932).) In other words, the memory based responses are heavily dependent on the similarity judgments. To account for this we would need a model of semantic structure which would allow similarity judgments to be computed from it (by a process model). One might think that a natural model would be something like a lexicon, including semantic features. D'Andrade notes that Shweder (1969) argues that such a model is not appropriate and that 'the basis on which the respondents make similarity judgments on this type is the degree to which attributes contiguously go together in making up a symbolic behavioral type, which he considers a special type of learned cultural construct.'

But given the results of the present study, this is somewhat paradoxical. If a 'behavioral type' is a learned construct, then presumably it is learned from people's behavior, that is, if attributes go together in a person's behavior, then they will go together in the type of which he is an instance, and thus the attributes will be semantically similar. But the results of D'Andrade's study show that it is precisely this that does not happen.

A possible way out of this bind is to suppose that people have predispositions to form behavioral types in certain ways, and that, in fact, those attributes which go together in people do not necessarily go together in the formation of a behavioral type. For example, an 'evaluation' component is a strong characteristic of personality attributes. That is, people tend to consider attributes good or bad. There may be a tendency to include only 'good' attributes or only 'bad' attributes in a behavioral type, even though this doesn't correspond to actual behavior. Saints and devils are rarer than the belief in them.

Actually there is a sense in which it may not be correct to describe what happens in these studies as 'forgetting.' Consider impression formation studies. In these experiments a subject is read a list of adjectives describing a person and then is asked to rate that person on a good-bad scale. It is well-known (Anderson, 1965; Chalmers, 1969) that in these judgments there is a primacy effect in the order of adjectives. That is, the adjectives presented first have more of an effect on the evaluation judgment than do later adjectives in the sequence. On the other hand, it has also been shown that in the memory for these adjectives there is mostly a recency effect. That is, the last adjectives in the sequence are remembered the best. Thus although subjects can remember the later adjectives better than the earlier ones, they don't use them in forming their judgments.

In this light we might consider the experiments that D'Andrade studied to be complicated impression formation studies. We conceive of the subject as having an impression of each person at each time and this impression changing with the added behaviors that he observes. Thus at the end of the experiment, when he is rating the other people, the subject might remember certain behaviors (the later ones) better, but might have his judgment more influenced by the earlier behaviors. (One could probably study this last proposition with existing data.) In this sense, forgetting does not seem to be a complete description of the process.

Geoghegan

Geoghegan's paper is an attempt to develop a theory of "the means by which human beings actually produce and interpret the message forms appropriate to a given form of verbal behavior." The paper illustrates many of the issues and controversies current in the cognitive sciences and thus is perhaps worthy of a somewhat longer discussion than most of the other papers. The particular setting is a study of the system of personal address in Samal. An attempt is made to develop a theory which will provide a cognitive representation of part of this system.

The bulk of the paper concerns a theory of the selection of a "name-type." This selection has been preceded (according to the theory) by a group of operations called an address form type selection routine. When the address form type that has been selected by this routine is N (for name), then the name-selection routine is invoked to elect the appropriate name. This routine consists of two operations. The first, the name-type rule, determines the type of personal name. A second operation then determines the lexical realization for this rule. It is the name-type rule that is studied in this paper.

There are seven possible name-types. Three of these have realizations as names of specific individuals. These types are represented as TN (true name), NN (nickname) and PN (pet name). Corresponding to each of these types is a proname type (labelled TN', NN', and PN', respectively) whose realization does not have the form of a personal name. An example is glossed 'old person.' The seventh type (labelled T) is an 'honorific' which does not correspond to any type of personal name.

The problem is to determine under what cognitive conditions a given name-type is used. According to theory, a "marking rule" is used to select the name-type. First, there is an "unmarked output," that is, a name that is used if none of the marking operators is applied (this is something like a "starting-state" in automata theory). The unmarked output can be either TN, NN, or PN. Which it is depends on the history of interaction between the addressor and addressee and is not given in the marking rule.

There are five "marking operators" which may apply to the unmarked output when the appropriate "marking cues" (attitudes, etc.) apply. They are a (positive affect), a' (negative affect), x (anger), d (deference) and n (addressee's name not known). Each of these operators maps a sub-set of outputs (names) into another sub-set. For example a (positive affect) will take NN into PN, and TN into NN. (We have not listed all the mappings that a performs.) In other words, if the unmarked output is NN and a applies, then the output will be PN. Likewise, if the unmarked output is TN and a applies, then the output will be NN. More than one operator can apply, so that if the unmarked output is NN, then if both a and d apply, a will take NN into PN and d will take PN into PN', so that the output of the marking rule will be PN'.

In short, the marking rule is an elegant theory of how various cognitive or emotional states determine the type of name that is used in an address situation. As such it is a valuable model of structure. What I would like to do is to discuss the status of this work as theory, with particular reference to work in related fields and to point out that models of structure do not necessarily yield models of process.

First, it is important to realize that the discussion that I have so far given of the marking rule actually does not state the entire theory. For there are restrictions on the combined use of marking operators that do not appear in that discussion. In other words, Geoghegan's Figure 2 does not contain the entire theory. For example, a and a' cannot apply together. (That is, both positive affect and negative affect cannot simultaneously apply.) Geoghegan discusses how such marking sequences can be eliminated on "structural grounds" (their use forms a cycle). But, as he points out, there are other combinations of operators, namely a' and d, which may not simultaneously apply but which cannot be eliminated on such structural grounds.

In order to account for these restrictions, Geoghegan lists (in Table 1) all 35 possible marking sequences, that is sequences of marking operators which it is actually permissible to apply. For example, a,d,a is a possible marking sequence, which first applies a, then d, then a again. But note that Table 1 does not specify the entire theory. For example, a,d,a may apply when NN is the unmarked output, yielding T as the output, but it may not apply when PN is the unmarked output, for after a applies, yielding T, according to Figure 2, d may not apply. Thus both Figure 2 and Table 1 together are needed to specify the theory.

This is unfortunate, for what appeared at first as a reasonably elegant and simple theory (Figure 2) now turns out to need a list of 35 possible marking sequences to be completely specified. So one attempts to construct a theory of Table 1, that is, to ask if a few simple rules added to Figure 2 wouldn't be sufficient. An adequate rule is that any marking sequence is permissible if 1) it leads (in Figure 2) from an unmarked output to a marked output and 2) if the sequence contains a or d then it cannot contain a' or x. This rule produces exactly the 35 marking sequences in Table 1. It is not clear from Geoghegan's discussion whether Table 2 was produced from a rule such as this or whether it was generated directly from some kind of data (i.e., informant interviewing). It seems reasonable to conclude that it could not be entirely generated by data, because there seems to be nothing (at first sight, at any rate) in an elicitation technique which would correspond to the order of the operators in the marking sequence. Thus, for example, both a,n and n,a are permissible marking sequences. This follows from the rule that has just been stated, but was there anything in the informant interviewing that would show this?

At this point it becomes relevant to ask, what kind of a theory is this "theory or marking rules?" In particular, to what extent is Geoghegan justified in calling the theory an "information processing (IP) routine?" To my mind, the answer to this question is, not at all.

A marking rule is no more an information processing routine for the encoding of address forms than a rule of grammar is an information processing routine for the encoding of sentences.

What is an information processing routine? It is a model which is supposed to represent the actual processes (at some level of abstractness) that a person goes through when he is performing some cognitive operation. In some ways the course of temporal events in the model (IP routine) is supposed to correspond to (be isomorphic to) the course of temporal events that is happening to the subject. Thus, for example, in the Newell and Simon (1963) GPS model, if the model formulates a sub-goal, this is intended to represent a human's forming of a sub-goal. I think that a survey of other theories that have been called information processing theories would find that this condition (isomorphism between human and model processing) was a condition that the authors intended to meet.

Now, to what extent is this condition true of the marking rule theory? Looking at Figure 2, one at first senses that here we have just such a theory. The seven outputs can be considered to be changes in states that take place over time. And in fact Geoghegan writes as if this is the case. For example, he writes:

> Suppose, for example, that Ego were to take NN as the unmarked output for a particular address situation. If he wanted to encode none of the cues available at this point in the process (those associated with the operators a,a',n,d, and x), then NN would remain in effect as the final output, and application of the rule would cease. If he wanted to encode one of the 'positive affect' cues, on the other hand, use of the operator a would occasion a shift in effective output from NN to PN (see Figure 2).

> At this stage of the process, several additional encoding options would be available. Ego could choose to encode no further information (with PN becoming the rule's final output), or he could continue the application by encoding information associated with either n, d or a.

But does Geoghegan really intend to say that the cognitive operations that a speaker of Samal goes through actually occur in the same order as the corresponding operations in the model? If so, what is the evidence for this?

Consider some cases. Suppose the marking sequence a,d is applied to the unmarked output NN. Are we to suppose that some stage the Samal speaker has (mentally) encoded a but not d? What is the evidence for such a conjecture? When he has encoded a, is he somehow thinking of the PN? What is the difference between applying the marking sequence a,d and the marking sequence d,a in the model supposed to correspond to in the cognitive operations of the speaker? Rather isn't it like the corresponding situation in syntax, where, (assuming we first develop S as NP+VP, i.e. Sentence as Noun Phrase plus Verb Phrase) we can then (in a context-free grammar, at any rate) develop either NP or VP next and it makes absolutely no difference to either the sentence or phrase marker that has been generated.

The reason that examples from current linguistic theory come to mind so readily is that the theoretical situation seems to be the same. A marking rule is much more like a rule of "competence" than a rule of "performance." As such it can bear no claim to the name "information processing system." The correspondence between the operations of the model and the operations that a human goes through is tenuous at best.

That this is true for linguistic theories of competence (that is, grammars) is, by now, well known. This is so not only in principle, but in fact. See Watt (1970) for an incisive critique and many references. It seems to be the case that sentences which are simple from a linguistic point of view are not simple from a psychological point of view. Of course, if your grammar is a model of competence rather than performance, then there is no reason that it should correspond to performance.

The reason that we have spent some time discussing the status of marking rules as theory lies in our search for structure and process. At first sight the existence of structure and process in this theory seem obvious, namely, Figure 2 with auxilliary assumptions represents both. But further consideration leads to the conclusion that the process model, if any, incorporated into this theory, is different in kind from most of the other processes discussed at the conference. It seems unlikely that people produce sentences by first dividing a notion S into a noun phrase followed by a verb phrase. Likewise it seems unlikely that people choose their forms of address in a manner indicated by the marking rule theory.

Geoghegan's work is a valuable description of the conditions under which various name-types are used. The above remarks are made simply to point out that by describing such conditions one does not automatically describe human processes.

Clark

In his paper Clark investigates how people understand negation. Although much of the work depends on the detailed specification of a process model, descriptions of structures on which the process model operates are also necessary. To get a feel for how the proposed process works, perhaps it is best to take an example which is slightly unnatural, that is, an example in which we can perhaps notice that we are calculating. Suppose I said, "John is not unreliable." To understand this sentence, I first note that unreliable means "not reliable" and then, because the "not" negates the sentence, the entire sentence means something like "John is reliable" (or perhaps "John is sort of reliable"). In other words, negative elements are processed serially, with each negative encountered changing the meaning of the core (non-negative) part of the sentence.

We will discuss the model for the comprehension of negatives with the following experiment in mind. A subject is presented a display containing a sentence and a picture and has to decide whether the sentence is true or false of the picture, pressing one button for "true" and another for "false." The time is recorded from when the display was presented until their response is made. The model attempts to account for these "verification latencies."

The pictures in the display contain, for example, a star above a plus, or a plus above a star. The sentence for this picture might be The star is above the plus, or The star isn't above the plus. The model proposes that in Stage 1, the subject represents the first sentence (using A and B appropriately) as (A above B) and the second sentence as (false (A above B)). If we assume that A is above B in the picture, then the first sentence is a True Positive and the second is a False Negative. Similarly, the False Positive is represented as (B above A) and the True Negative as (false (B above A)). Thus, by a representational process, the subject arrives at a structure for the sentence. The next process (Stage 2) represents the picture as (A above B).

Stage 3 compares the representations of sentence and picture obtained in the first 2 stages and calculates a truth index. This index starts out with the value true in it, and the value changes whenever a mismatch occurs between the sentence and picture representations. In Stage 4 the subject responds with the final value of the index calculated in Stage 3.

The Stage 3 comparison process works by comparing first the embedded strings of sentence and picture, changing the truth value if they do not match. It then compares the embedding strings of sentence and picture, once again changing the truth value if they don't match.

By making a few assumptions about the times needed to complete each of the processes, Clark predicts the verification latencies in this experiment quite well. The details are too complex to go into here, and the paper makes a number of important points. Among these are the claim that 'negation is fundamentally a semantic notion' and that presupposition and scope of negation play a role in the process of comprehension of negatives.

To summarize, the structural model is a representation of sentences (and of pictures) which contains 'embedded' and 'embedding' strings. The process model is given as a branching diagram or flow-chart which calculates a 'true' or 'false' response depending on the structure of the sentence and picture.

I think it is worth pointing out that Clark's study is the only one in this collection to construct process models which analyze units larger than a word (Crothers analyzes the structure of paragraphs). In proposing precise models of how not combines with other words to form compounded meanings, Clark goes beyond most of the work in the field.

Crothers

Crothers' paper is an attempt to find ways of characterizing the semantic structure of paragraphs. The system that is used defines itself by its details, and thus it would be impractical to try to explicate too much of that system here. Rather, I will for the most part content myself with a description of the kind of thing that it is, and the kinds of things that it isn't.

First, it is a model of structure, and, as such, is independent of any particular process or experiment. Thus it is not a model of comprehension or memory. Rather, according to Crothers, a model of structure should be developed prior to any such model of process. Thus, in terms of the well-known distinction, the structural model is a model of competence, rather than one of performance.

Crothers' approach is to analyze a paragraph so as to exhibit the intuitive semantic relationships. He does not start with a semantic theory, but rather invents elements as needed to satisfy his intuition. Crothers calls this method 'inductive generalization.'

Of course, it is well known that proper induction demands more than simply the noting of a feature in a particular observation and then stating that this feature is true in the general case. For there will always be features true of any particular situation that do not generalize. Rather, as Peirce (1957) pointed out, one must state what one expects to find before the observation is made. If, as more and more paragraphs are analyzed, new elements keep appearing, then this difficulty will apply, and the method of inductive generalization won't work. But suppose, on the other hand, that after the analysis of a number of paragraphs, a stock of elements has been built up and these elements are then sufficient to explicate all paragraphs that follow. In this case the method of induction has been properly applied.

Crothers analyzes 3 paragraphs in this paper. Each analysis introduces new elements into the theory. Thus there is no inductive evidence that a correct theory has been found. Of course, as Crothers states, this is the beginning of work on paragraph structure and only by working through a number of examples can one hope to arrive at a general theory. Thus only the future will tell whether the approach is successful. At the point when new elements do not have to be invented for each new paragraph, an adequate theory will exist.

On the other hand, one might take a different approach. He could say that the theory does not consist of all the detailed elements. Rather, the hypothesized structure of the paragraph constitutes the theory and thus is the subject of the inductive generalization. The theory then would be something like the structure of paragraphs may be represented as a network. But at this point the theory is almost vacuous since networks are a powerful means of representation and could represent many different proposals. The reasoning is similar to Chomsky's (1970) argument against Lakoff's (1969a,b) position on generative semantics. Lakoff claims that semantic representations are, in fact, syntactic phrase-markers, and thus no rules of interpretation are needed for semantics, that is, no rules to map from the phrase-markers into semantic representations. Chomsky argues that ". . . virtually any proposal that has been made concerning semantic representation can, in equally uninteresting ways, be reformulated so as to use phrase markers for semantic representation . . . . It is difficult to imagine any coherent characterization of semantic content that cannot be translated into some 'canonical notation' modeled on familiar logics . . .". Since networks are graphs and thus contain trees as a sub-set and since phrase-markers are trees, it follows that networks are more powerful than phrase-markers. Thus the above argument follows even more strongly for networks. Saying that a network is the proper representation for semantics is not saying very much. Thus the detailed elements must be part of the theory and confirmation awaits the analysis of future examples.

Here I will just briefly review Crothers' method for analysis of paragraphs. First, each sentence is decomposed into 'basic sentences.' For example, the sentence 'steel is made by combining iron and carbon' is broken down into the 2 basic sentences 'steel is made' and 'iron and carbon are combined.'

The second step is to establish a 'semantic hierarchy' on the basic sentences. The semantic hierarchy depends on knowledge of the language but is almost independent of the particular paragraph. That is, the paragraph determines the set of basic sentences, but once that is given, the semantic hierarchy is set. The hierarchy is determined by implication and implication-like relationships. Thus, for example, 'The writing instrument was lost' dominates 'The pen was lost.'

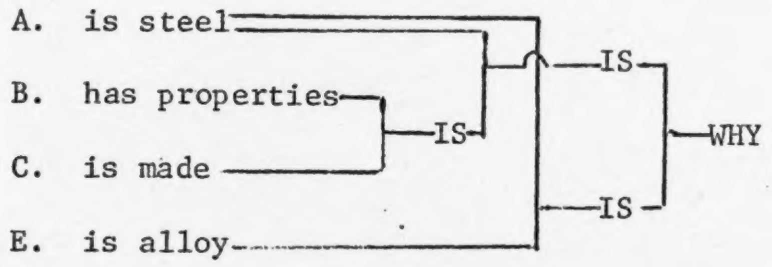
The final step in the analysis is to determine the 'fundamental structure' of the paragraph. The nodes from the semantic hierarchy are still there but there are many more nodes and connections made between nodes. These connections depend on the paragraph and thus are not facts of general knowledge as is the case with the original semantic hierarchy connections. The connections are labelled, each label corresponding to a logical relationship. For example, one label is 'OR.' Also, correspondences between nodes have to be enumerated. This is done in a straight-forward manner which is not part of the network formalism, but it could be, as it was in an earlier version of the paper.

This description of Crothers' method is, of course, far too sketchy to give the reader an understanding of it. But the method is best described by examples, and thus the paper should be worked through. Crothers gives many examples and, although the system is somewhat complex, the writing is clear. One point that should be kept in mind was stated above, namely the difference between the semantic hierarchy and the fundamental structure. The hierarchy represents known implication relationships and does not depend on the facts stated in the paragraph. The fundamental structure, on the other hand, does state relationships expressed in the paragraph.

Of course the above procedure is not formalized to the point where explicit rules can be given so that the result can be automatically generated. But, as Crothers points out, if we want to study psychological processes that depend upon the structure of text, then we need some way to calculate that structure and human intuition is by far the best method available at this time. Crothers' method may be taken in this spirit as a guide to intuition and a standardization of its results.

One question that we may ask is, why choose the network formalism for the semantic representation? An answer is that this formalism is familiar from existing work. For example Quillian's (1968) program has the same structure, namely nodes with labelled connections (representing relations) between them.

There appear to be other equivalent modes of representation, however. For example, can the predicate calculus express the same relationships? Here is Crother's sub-graph (5).



It appears that the following predicate calculus formula is in some sense equivalent (where 'WHY' is a 2-place predicate, 'ALLOY' is a 1-place predicate, and 'HAS PROPERTIES AND IS MADE' is a 1-place predicate):

WHY (ALLOY (steel), HAS PROPERTIES AND IS MADE (steel)).

A translation of this formula might be, 'Steel has properties and is made because steel is an alloy.' Of course, this is the 'higher' predicate calculus, that is predicates may be arguments of other predicates. But it is well known that this is necessary for an adequate description of English. (For a philosophical treatment, see Reichenbach (1947) and for a linguistic approach see Weinreich (1965)).

One argument that might support the use of the network notation instead of a predicate calculus formulation is that the network makes clear such concepts as 'cycles' and 'paths.' Thus it might be more natural for psychological work than the logical notation, whose formulas do not so readily yield their structure.

Burton and Nerlove

Burton and Nerlove's paper is unlike the others in that it is primarily methodological. The problem they confront is that of doing a triads test of judged similarity. (A triads test here is one in which a subject is asked to select the one item of 3 presented that is most different from the other 2). When the number of items becomes large, then the number of different triads becomes so large that it is impractical to test all the triads.

The solution that Burton and Nerlove present for this problem is the recommendation to do a 'balanced incomplete block design.' This is a design in which a subset of the triads is chosen so that each pair of items appears in exactly λ triads. For small λ this can greatly reduce the number of triads. There is no algorithm known for generating such designs in general. Burton and Nerlove offer solutions for a number of special cases.

In order to test how adequately the balanced incomplete block designs do in capturing the structure obtained from a complete triads test, a set of triads data on vegetable terms was collected. In addition to a complete triads test, various incomplete balanced block designs were run. Random subsets of triads were also tested.

The data from each of these tests were analyzed by a multidimensional scaling program. The point was to see how similar were the pictures obtained, on the one hand from the complete design, and on the other, from the incomplete design. In order to do this, rank-order correlations on the interpoint distances were computed between the complete test and each of the other designs.

One result was that each of the incomplete balanced block designs had a higher correlation with the complete test than did the random subset of the same number of triads. In other words, block designs produce results more similar to complete designs than do random subsets. A second result is that, as λ increases, so do the correlations. In other words, increasing the number of triads helps. A third result is that for values of λ greater than 1, incomplete designs do quite well. A second experiment, on kinship terms, also supported these results.

The general method employed by Burton and Nerlove is applicable to any scheme of analysis of triads which first reduces the data to a similarity measure. Thus, for example, hierarchical models can be studied by putting the data through a Johnson (1967) cluster analysis, and comparing the incomplete and complete designs. However, if the data are to be analyzed by a method which works directly on the triads, then these designs might not be applicable.

Greeno

Greeno's paper contains a report of 3 studies which deal 'with some aspect of the process by which a subject uses relationships between elements of his cognitive structure.' The studies aren't integrated by a coherent theory, but were designed to acquire information which might lead toward such a theory. As such, it is perhaps most appropriate to deal with each experiment on its own terms.

The first experiment was designed to study a problem which involved transformation of information but did not involve calculation of an answer (in the usual sense). One group of subjects (the conceptual group) first memorized 4 formulas, for example 'driving time = arrival time - leaving time' or 'distance = driving time x average speed.' Then (the main task), the subject was asked to determine whether the value of a given variable could be determined if he were given values of certain other variables. An example is

Arrival time--gas used, gas mileage, average speed, leaving time. In this problem the subject had to decide whether he could calculate 'arrival time' given the other 4 variables. The second (nonconceptual) group had the identical task except that instead of concepts like 'driving time' being used in the materials, letters were substituted. Thus a formula they would memorize would be V = F - L. Their problems would also be stated in terms of letters, for example

F--G,M,A,L,

meaning, can F be calculated if G,M,A, and L are given?

Response times were measured for each of the problems and results were as expected. The more variables given the longer the response time. Also, the conceptual group had shorter response times than the nonconceptual group. An interesting result was that problems with positive answers took about the same time for both groups of subjects, but problems with negative answers took much more time for the nonconceptual group than for the conceptual group.

These results lead Greeno to speculation concerning models for structure and process. First consider the conceptual subject. He might have list structures of the form V,F-1 or D,V·A . To answer a problem he would apply a sequential decision process, corresponding to a decision tree. The important point is that in order to come up with a negative answer all the possibilities have to be exhausted. Thus negative answers would take longer than positive ones, which conforms to the experimental result for nonconceptual problems.

For conceptual problems Greeno suggests that the structure be considered (instead of a list) a 'relational net' with the concepts represented by arcs and the relations or operators represented by nodes. The notion here seems to be that the concepts are integrated in one network. Instead of appearing independently in a number of different formulas, each concept appears once and is related through the net to all other concepts.

A positive answer can be given by following a process which finds paths through the network. Although Greeno does not specify how the model decides on a negative answer, he intuitively feels that the difference in response times for positive and negative answers based on the relational network would be much smaller than the difference based on the decision tree.

Passing by Greeno's second study (on learning to apply the binomial formula), we will briefly discuss his third experiment, which studied processes of deductive reasoning. The method was to give a subject a premise and a 'conclusion.' The subject had to decide whether the conclusion followed from the premise, whether the negation of the conclusion followed from the premise or whether neither the negation nor the conclusion followed from the premise.

In this way Greeno discovered that many of the rules of formal logic are not followed. One example should suffice, especially since it is a problem that has occurred in countless introductory logic classes. The rules of material implication state that 'not p' implies 'if p, then q.' Subjects will not accept this implication.

Greeno's conclusions are that subjects judge the derivations as if they were interpreting real events. For example, he writes, "'if, then' statements are usually intended to describe conditions in which the antecedent applies. Since this is the usual interpretation of the sentences it is counterintuitive for a subject to deduce an 'if, then' statement from the negation of its antecedent."

Greeno points out the importance of the study of cases such as this, that is, where what is important is 'the criterion of usefulness in interpreting events that occur.' Note that this is not only a matter of considering semantics in addition to syntax, for the subject 'inconsistencies' will still be there when semantical considerations are added. That is, adding a semantic model to a logical system does not make the derivations non-logical. But there is now some sociolinguistic work which speaks to this general question. I am thinking, for example, of Sacks' (1972) studies of conversation, in which this notion of processes used by speakers is studied. Of course, Ludwig Wittgenstein also had much to say concerning this issue.

General Discussion and Summary

What sort of picture of human cognition emerges from the groups of papers discussed here? The first answer is that both structure and process are essential elements of cognition, and that they are not independent, but rather, intimately related. All the papers may be interpreted as talking about processes that apply to structure. On the other hand, this framework may be still too simple. The interrelations between process and structure may be quite a bit more complex than the 'structure precedes process' viewpoint taken here. It is possible that often structures will have to be considered as arriving from processes. To take one example, Clark's representation of the structure of negative sentences would of course have to be arrived at by a process, starting, presumably, with the surface form of the sentence. Of course, this surface sentence would itself be represented by a structure. Thus we might expect complicated relations of structure and process to be necessary in our models.

The second reply to the question about a general view of human cognition is that, unfortunately, aside from the very abstract level of structure and process, one can't be determined. The interrelations between the structures and processes studied in each of the papers have not been worked out. Each paper analyzes a limited range of phenomena. A number of different kinds of structures are proposed, for example, Euclidean spaces, lists, trees and networks. There is not one view which at the moment can cover most of the phenomena.

There is one other integrating idea, perhaps as general as structure and process, which is beginning to appear in this work. This is an acknowledgment of the relational character of knowledge, and the development of ways to study that knowledge. While the collection of papers presented at this conference represents a beginning of a new kind of work in cognitive psychology, it also in some ways represents an end, in fact, an end of innocence. Consider, for example, the paper by Rumelhart and Abrahamson. Words have been scaled into multidimensional space, and analogies solved in that space. There is no acknowledgment of the relational character of many words. Of course, the paper deals only with nouns, most of which may be considered non-relational. But that, too, is symptomatic. If one concentrates on nouns, relational aspects of meaning may be minimized. Perhaps the reason that psychologists have turned to nouns when they want to study linguistic materials is to be found in the difficulty of dealing with relations. But many of the papers under discussion deal with relations.

Of course, linguists, anthropologists, and, especially philosophers and logicians have long recognized the central role that relations play in language (though even in anthropology the relational aspect of kinship terms has not always been recognized). (In fact, Filmore's work on underlying semantic relations was often mentioned in the discussions.) That in fact an end of innocence is here is clear from a consideration of more recent papers by Rumelhart, Lindsay and Norman (1970) and Kintsch (1971). In these papers, not only is the relational character of knowledge acknowledged, but it is systematically analyzed, drawing heavily upon the classical logical notions. Although the great bulk of psychologists still go about their business not caring about these elementary distinctions, it is clear that these papers (and others) will be quite influential and that non-relational objects have had their day (and say).

Another observation is the great concern with semantics. All of the papers considered here (with the possible exceptions of Greeno's and Geoghegan's which might, using the usual unclear distinction, be classified as 'pragmatic') are directly concerned with semantics. Semantics is providing a setting for much of the work in cognition.

Work in human cognition is proceeding, and while we do not yet have anything like an adequate general theory of cognition (or semantics, for that matter), we can hold out hope that the future will provide greater insights.

REFERENCES

Anderson, N. H. Primacy effects in a generalized order effect paradigm. Journal of Personality and Social Psychology, 1965, 2, 1-9.

Bartlett, F. C. Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press, 1932.

Bendix, E. H. Componential analysis of general vocabulary: the semantic structure of a set of verbs in English, Hindi and Japanese. Bloomington: Indiana University, 1966. Also International Journal of American Linguistics, 32, part 2.

Borgatta, E. G., Cottrell, L. S., and Mann, J. H. The spectrum of individual interaction characteristics: An inter-dimensional analysis. Psychological Reports, 1959, 56, 279-319.

Chalmers, D. K. Meanings, impressions and attitudes: a model of the evaluation process. Psychological Review, 1969, 76, 450-460.

Chomsky, N. Some empirical issues in the theory of transformational grammar. Mimeographed, Indiana University Linguistics Club, November 1970.

Evans, T. G. A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (ed.), Semantic Information Processing. Cambridge: MIT Press, 1968.

Henly, N. M. A psychological study of the semantics of animal terms. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 176-184.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-253.

Kintsch, W. Notes on the semantic structure of memory. University of Colorado, Department of Psychology, Report CLIPR-4. 1971, 1-90.

Lakoff, G. On derivational constraints. In R. I. Binnick et al (eds.), Papers from the Fifth Regional Meeting. Chicago Linguistic Society, 1969a.

Lakoff, G. On generative semantics, mimeographed, 1969b.

Mann, R. D. The relation between personality characteristics and individual performance in small groups. Ph.D. dissertation, University of Michigan, 1959.

Miller, G. A. A psychological method to investigate verbal concepts. Journal of Mathematical Psychology, 1969, 6, 169-191.

Newell, A., and Simon, H. A. GPS, a program that simulates human thought. In E. Feigenbaum and J. Feldman (eds.), Computers and Thought. New York: McGraw-Hill, 1963.

Peirce, C.S. The order of nature. In Tomas, V., (ed.), Peirce's Essays in the Philosophy of Science. New York: Liberal Arts Press, 1957.

Quillian, M. R. Semantic memory. In M. Minsky (ed.), Semantic Information Processing. Cambridge, Mass.: MIT Press, 1968, 216-270.

Reichenbach, H. Elements of symbolic logic. New York: Macmillan Company, 1947.

Rumelhart, D. E., Lindsay, R. H. and Norman, D.A. A process model for long-term memory. Mimeographed, University of California, San Diego, 1970.

Sacks, H. Aspects of Sequential Order in Conversation (to appear).

Shweder, R. A. Symbolic arrangement as a measure of interpersonal behavior variables. Unpublished, 1969.

Watt, W. C. On two hypotheses concerning psycholinguistics. In J. R. Hayes (ed.), Cognition and the development of language. New York: Wiley, 1970.

Weinreich, U. Explorations in semantic theory. In T. A. Sebeok (ed.), Current Trends in Linguistics, V. III. The Hague. 1966.

Whitehead, A. N. Process and reality. New York: Free Press, 1929.

THE USE OF THE BALANCED BLOCK DESIGN FOR THE TRIADS TEST OF JUDGED SIMILARITY[1]

Michael L. Burton and Sara B. Nerlove

University of California, Irvine

---

[1] This work was partially supported by a National Institutes of Health Grant, PHS MH 17468-01, to A. Kimball Romney. The authors are especially grateful to Marc Miller for spending many hours of assistance in the computer work for this paper and for his helpful comments. We also want to thank A. Kimball Romney and Volney Stefflre for their valuable suggestions.

Abstract

The use of balanced incomplete block designs as a means for extending the use of the triads test of judged similarities to large semantic domains is examined. Balanced incomplete block designs are applied to triads data from the semantic domains of vegetables and kinship and the resulting multidimensional scaling configurations indicate that such designs provide an adequate picture of the structure of the domain; however, more than one balanced design should be used in gathering data and each design should be more inclusive than a minimal subset of these data. Procedures for obtaining solutions for different *n*s between 9 and 25 and for different sized subsets of a given *n* are presented.

The aim of this study is to investigate the possibility of extending the use of a judged similarities method important in exploring the structure of semantic domains (Burton, 1968; Romney & D'Andrade, 1964; and Wexler & Romney, 1969); and, at the same time, to obtain some information about the nature of similarity data. This method of judged similarity is the triads test, a sorting test devised by Romney and D'Andrade (1964) based on procedures used in clinical psychology (Kelly, 1955) and in psychophysical measurement (Torgerson, 1958). The data from this test can be used to uncover the dimensionality of meaning of the domain to which the stimuli presented in the test belong.

The triads test consists of presenting sets of three terms to the subject who is instructed to choose the one of these terms that he regards as most different in meaning. For example, one triad in the domain of dinner vegetables is peas, lima beans, and cauliflower. (On the basis of the dimensions of various attributes of the objects such as color, size, and shape, it might be predicted that in this case a majority of subjects would choose cauliflower as most different leaving peas and lima beans as a pair in this context.)

A salient feature of this method is the relative simplicity of the judgments required of the subject. Not only can such a method be used cross-culturally, but also it can be used with nonliterates. To use the method is a matter of producing a sentence frame consisting of either the question "which one is most different" or "which two are more similar" and putting any given set of three items into the substitution slots of that frame.

From each triad, information pertaining to three pairs of items can be obtained. Thus the choice of the most different item in the triad A, B, C provides a measure of the similarity of A and B to each other relative to C; and, correspondingly, it also provides a measure of similarity for the pairs A,C and B,C. In a complete triads test, that is, one in which all possible sets of three are presented, each pair occurs with each of the other $n-2$ items. For a given pair of items, it is possible to compute the number of contexts in which two members of that pair remain together. The similarity of any two items, then, is the total number of times these two items are paired together divided by the number of subjects.

The number of triads increases proportionally to the fourth power of n. Therefore, for eight items there are 56 triads in a complete test, for ten items there are 105 triads, and for 15 items there are 455. Thus for a domain with a large number of items, the triads test is impractical.

As Coombs (1964) has pointed out, some methods have great capacity and hence great redundancy as compared to the method of paired comparisons. For example with a full complement of pairs, each pair appears only one time, while each item occurs $n-1$ times. With a full complement of triads, however, each pair appears $n-2$ times, while each item occurs $(n-1)(n-2)/2$ times. To present stimuli in sets of three (triads), four or more items is to use methods of increasingly greater redundancy. The potentially large number of presentations of each pair can lead to subject fatigue and boredom, which could cancel the value of using the more powerful method. As a solution to this problem, Coombs suggests

> ...we might take advantage of the power of some of these methods by using a subset of the total possible presentations. If we may obtain more information from a subject on a single presentation without taxing him, fewer presentations are called for to obtain the same amount of information than are called for by a less powerful method. (Coombs, 1964, p. 42.)

Since the similarity measure derived from the triads test is based on the mean number of times that two items have been paired together, we have chosen to investigate subsets of items for which each pair of items appears equally often. These are known as balanced incomplete block designs (Cochran & Cox, 1957; Durbin, 1951; Gulliksen & Tucker, 1961; Yates, 1936.)

For balanced incomplete block designs, λ is used to denote the number of triads in which each pair of items occurs. Thus, the minimal balanced design has $\lambda = 1$, and the complete triads test has $\lambda = n-2$. There are $n(n-1)(n-2)/6$ triads in a complete test. The number of triads in a balanced incomplete block design is obtained by multiplying the number of triads in a complete test by $\lambda/n-2$ or $\lambda n(n-1)/6$. Two other conditions are necessary for a balanced incomplete block design of triads:

(1) $rn = 3b$

(2) $\lambda = 2r/n-1$,

where r = the number of replications of each stimulus and b = the number of triads in the incomplete block design.

In this paper designs for various λ for $n = 9$ and for $n = 15$ are used. Designs have also been presented for various ns between 9 and 25 and for different values of λ for these ns (see Appendix). (We have generated most of the designs which are in both the body of the paper and in the appendix ourselves and would appreciate any information about places where these designs might already have appeared in print.)

Method

The domain of vegetables was examined in order to compare the relative stability of the structure of various triad subsets to the structure of all possible triads. For an initial examination of this problem, the approach was to work with a set of stimuli small enough to be administered in its entirety of combinations to each subject. From these data, it was then possible to carve out various subsets from the total number of triads, keeping the sample constant and having the actual structure of the full complement for comparison.

Defining the Domain

Two tasks were used to elicit the domain. One task was a free listing of all vegetables a subject could think of. The second task was a sorting task for which no limits upon size or number of piles were imposed. A deck of cards with the names of all the more common vegetables which the investigators could recall supplemented by a basic American cookbook index was sorted into piles. The subject was interviewed as to the basis of his classifications. It was clear from the sorting task that dinner vegetables were distinct from salad vegetables. Then from the free listings, the nine most frequently occurring dinner vegetables were chosen and each one of these was assigned a number.

Experiment 1

Subjects and Stimuli

One hundred male and female undergraduate students at the University of California at Irvine were presented with the 84 possible triads of the nine vegetables in a random order. The vegetables are as follows: 1. turnips, 2. potatoes, 3. peas, 4. corn, 5. lima beans, 6. spinach, 7. cauliflower, 8. onions, 9. asparagus.

Choosing Designs for Analysis

The complete set, balanced subsets, split-halves of subjects, and random subsets of these triads data were analyzed by Shepard-Kruskal nonmetric multidimensional scaling methods, using the TORSCA program. (Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b; Young and Torgerson, 1968).

Generation of balanced incomplete block designs, n = 9.

Seven balanced subsets of 12 triads each for which λ = 1 were examined. In these designs n stimuli are presented in s sets of p at a time so that each stimulus occurs in r of the sets. Specifically nine stimuli are presented in 12 sets of three at a time so that each stimulus occurs in each of the four replication blocks.

The first design, design A, is one listed by Schucker (1959):

Rep I		Rep II		Rep III		Rep IV	
Set		Set		Set		Set	
(1)	1,5,9	(4)	2,6,9	(7)	3,7,9	(10)	4,8,9
(2)	2,3,8	(5)	1,3,4	(8)	2,4,5	(11)	3,5,6
(3)	4,6,7	(6)	5,7,8	(9)	1,6,8	(12)	1,2,7

(Schucker, 1959, p. 274)

Design B is one from Cochrane and Cox (1957):

Rep I		Rep II		Rep III		Rep IV	
Set		Set		Set		Set	
(1)	1,2,3	(4)	1,4,7	(7)	1,5,9	(10)	1,6,8
(2)	4,5,6	(5)	2,5,8	(8)	2,6,7	(11)	2,4,9
(3)	7,8,9	(6)	3,6,9	(9)	3,4,8	(12)	3,5,7

(Cochrane & Cox, 1957, p. 376)

The third design, C, was constructed according to the same principles as the Cochrane and Cox design. (For an example of generation of designs.

according to such principles in using a cyclical procedure see the discussion for n = 15 below; this procedure can only be used where n = 9 or n = 15.) Design C shares six triads with design B, but is completely independent of design A:

Set	Rep I	Set	Rep II	Set	Rep III	Set	Rep IV
(1)	1,2,3	(4)	1,4,7	(7)	1,6,9	(10)	1,5,8
(2)	4,5,6	(5)	2,5,9	(8)	2,4,8	(11)	2,6,7
(3)	7,8,9	(6)	3,6,8	(9)	3,5,7	(12)	3,4,9

Design D was produced by adding three to each number in design A. Thus the triad (1,5,9) became the triad (4,8,3). Design E is another design produced according to the same cyclical procedure as design C:

Set	Rep I	Set	Rep II	Set	Rep III	Set	Rep IV
(1)	1,2,3	(4)	1,6,8	(7)	1,4,7	(10)	1,5,9
(2)	4,5,6	(5)	2,5,7	(8)	2,6,9	(11)	2,4,8
(3)	7,8,9	(6)	3,4,9	(9)	3,5,8	(12)	3,6,7

Design F was produced by adding two to all numbers in design B. Design G was produced by subtracting two from all numbers in design A.

These seven designs are not mutually exclusive of each other, though each design is at least 50 per cent independent of any other design. Table 1 shows the degrees of overlap (number of common triads out of a possible maximum of 12) between pairs of designs.

Insert Table 1 about here

out of 21 pairs of designs, there are five with no overlap, six which overlap in one triad, seven which overlap in two or three triads and three which overlap in six triads.

Five designs of 24 triads each for which λ = 2 were examined. Design H was produced by adding together designs A and C. Design I was produced by adding two to all of the numbers in design H. Design J was produced by adding four to all the numbers in design H. Design K was produced by subtracting four from all of the numbers in design H. Design L was produced by combining designs F and G. Table 2 shows the numbers of overlapping triads between pairs of designs for λ = 2. None of the λ = 2 subsets are totally independent of any other, the overlap between them ranging from 8 to 58 per cent.

Insert Table 2 about here

Here the degree of independence among designs is not as large as it is for λ = 1.

There are three designs of 36 triads each for which λ = 3. They are produced by a system of cycles. Four cycles of nine triads each will produce the 36 triads which are needed for a balanced design where λ = 3. Each cycle is indicated below by listing the first triad and the last triad. The second triad is produced by adding one to each of the numbers in the first triad. Thus the triad succeeding 1,2,4, is 2,3,5. The entire cycle is generated by repeating the procedure nine times. Since this is base nine arithmetic, adding one to nine produces one. The four cycles for the first design, M, are

Cycle I   1,2,4 add one until 9,1,3;

Cycle II   1,3,4 add one until 9,2,3;

Cycle III 1,2,6 add one until 9,1,5;

Cycle IV 1,3,6 add one until 9,2,5.

It is easy to show that the above set of four cycles comprise a balanced design. Every pair of items must be included in exactly three triads. A pair may be separated by a numerical difference of one, two, three, or four. Since we are working in base nine arithmetic, a space of five is equivalent to a space of four (8 to 4 can be traversed by adding 5 to 8 or by subtracting 4 from 8). A balanced design must include each of these numerical differences in exactly three places. For example, the cycle 1,2,4...9,1,3 covers spaces of one (1 to 2), two (2 to 4), and three (1 to 4). When all nine triads are produced by cycling from 1,2,4 to 9,1,3; all pairs which are separated by one, two, or three spaces will have been included in one triad each. It can easily be verified that the spaces within the cycles for design M are such that each of the four spacings is included exactly three times.

The second design, N, is

Cycle I 1,2,4 add one until 9,1,3

Cycle II 1,3,5 add one until 9,2,4

Cycle III 1,4,5 add one until 9,3,4

Cycle IV 1,2,5 add one until 9,1,4

Cycle I of designs M and N are identical and thus they have a 25 per cent overlap with each other.

The third design, P, is

Cycle I 1,5,6 add one until 9,4,5

Cycle II 1,4,6 add one until 9,3,5

Cycle III 1,2,4 add one until 9,1,3

Cycle IV 1,3,4 add one until 9,2,3

design P shares two cycles with design M and one with design N.

There are also three designs for λ = 4: Q, R, and S. These designs are simply the respective complements of designs M, N, and P; i.e., design Q includes all of the triads not a part of design M and so on.

Split-half designs. Subjects who had been arbitrarily numbered from 1 to 100 were split into two groups (subjects 1 to 50 forming one group and subjects 51 to 100 forming the other) and the multidimensional scaling configuration of the data on all 84 triads for one half of the subjects was compared to that for the other half of the subjects. To ensure that the subjects had not been split along some line of natural clustering of differences, they were split in half into an even- and an odd-numbered group and the configuration of the data on all 84 triads for the even-numbered subjects was compared to that for the odd-numbered subjects.

Randomly chosen sets. Finally, three randomly chosen subsets of 12, 24, and 36 triads, respectively, were analyzed.

Results of Experiment 1

2nd Order

Figure 1 illustrates that for vegetables there is a large decrease in stress from one to two dimensions and a relatively small decrease in stress from two to three dimensions. This evidence indicates that a two-dimensional scaling representation of the domain is adequate.

Insert Fig. 1 about here

A program was developed that could produce from the basic body of data similarities tables for any subset either of subjects or triads (balanced or randomly chosen). The overall structure (all subjects, all triads) of

The domain that emerged from the multidimensional scaling analysis showed three tight clusters which might be labeled "roots," "leaves and stalks," and "seeds." Then Johnson's hierarchical clustering program (Johnson, 1967) was used to verify this feature of the scaling configuration (see Figure 2). The results were the same using both the Diameter and the Connectedness methods. Figure 3 shows a graphic plot of the vegetables in two dimensions, with the major groupings from the hierarchical clustering analysis indicated by dotted outlines.

Insert Figs. 2 and 3 about here

Having settled on a two-dimensional representation for the vegetables, we then did scalings for the seven designs for λ = 1, the five designs for λ = 2, the three designs for λ = 4, three randomly chosen sets, and two different pairs of scalings for which the 100 subjects were split into two groups of 50 each.

Given these two-dimensional configurations and their associated tables of interpoint distances, it was possible to compute rank-order correlations on the distances for pairs of configurations. These rank-order correlations are listed in Table 3 (for λ = 1), Table 4 (for λ = 2), Table 5 (for λ = 3), Table 6 (for λ = 4), Table 7 (for randomly chosen sets), and Table 8 (for split-halves of subjects).

Insert Tables 3 through 8 about here

Table 3, for λ = 1, shows that the design which produces a structure most similar to the scaling configuration for the complete triads is design C (rank correlation = .952). Correspondingly, the "worst" design is design F (rank correlation = .570). Since the random set of 12 triads (see Table 7) produced a configuration with rank correlation of only .402 to the configuration for complete triads, it can be concluded that the "worst" balanced block design is more reliable than a merely random set of 12 triads. In Figure 4, designs C and F are plotted along with the complete triads configuration.

Insert Figure 4 about here

One important reason why the rank correlation for design F is so low is that one of the 12 triads is *corn, peas,* and *lima beans.* These three items form a tight cluster in the overall structure. The three pairs *corn-peas, corn-lima beans,* and *peas-lima beans* are all contained in this triad. If these pairs had appeared in triads in which they were contrasted with items outside the cluster, e.g., *corn, peas, asparagus,* then the outside item would almost always have been judged as most different. Thus a similarity measure for *corn* and *peas* based on that triad would have been relatively high, whereas the similarity measure for the same pair based on the triad *corn, peas, lima beans* must be lower for the contrast is with a member of the same cluster. However, the subjects' choices were not evenly distributed among the three members of the triad. The distribution of choices of *most different* was as follows:

peas	13
lima beans	19
corn	68

The fact that subjects tended to select *corn* as most different had the effect of producing very low similarity measures for *corn-lima beans* and for *corn-peas*, and of placing *corn* on the opposite side of the diagram from *peas* and *lima beans*. When design F is combined with design G to produce the λ = 2 design, L, this triad also appears; but along with it are other triads (*potato*, *corn*, *lima beans*; and *turnip*, *peas*, *corn*) which emphasize the similarity of corn to peas and lima beans, thus reintegrating the cluster.

Table 4, for λ = 2, shows that the design which produces a structure most similar to the scaling configuration for the complete triads is design I (rank correlation = .898). Correspondingly, the "worst" design is L (rank correlation = .764). A random set containing 24 triads (see Table 7) produced a configuration with rank correlation of only .590 to the configuration for complete triads. Again, the evidence is that the "worst" balanced incomplete block design is more reliable than a merely random set of the same number of triads. In Figure 5, designs I and L are plotted along with the complete triads configuration.

Insert Figure 5 about here

Tables 5 and 6, for λ = 3 and λ = 4 respectively, show that the scaling configurations produced by these designs have high correlations to that for the complete triads. It would seem that a representation of roughly half of the triads is as reliable as a representation of half of the subjects (cf. Tables 5 and 6 with Table 8). Again, the "worst" balanced block design for λ = 3, design N (rank correlation = .945), is more reliable than a merely random set of 36 triads (rank correlation = .808) (Table 7). All three designs for λ = 3 have very similar rank correlations to the scaling configuration for the complete triads. Plotted in Figure 6 is one of the balanced designs for λ = 3, along with the complete triads configuration.

Insert Figure 6 about here

Figures 7 and 8 are graphs of the degree of overlap between pairs of designs against the rank order correlations for the same pairs of designs.

Insert Figures 7 and 8 about here

Figure 7, for λ = 3, shows that for 0, 1, 2, or 3 triads overlapping there is a great amount of variability in rank order correlation, but that for 6 overlaps this variability has decreased; and the rank order correlation has settled onto a relatively high value. Figure 8, for λ = 2, shows much less variability in rank order correlation at any degree of overlap, indicating that there is a much higher lowest bound on rank correlation for λ = 2. From the evidence about λ = 2 and the earlier discussion of the distortion which occurred in design F, it can be deduced that certain extreme events which cause very low rank correlations for λ = 3 have become much rarer for λ = 2. For the vegetables which have clusters of three members each, such events cannot occur for λ = 2 for they depend upon both the presence of a triad which has as its members three elements of a (three-element) cluster and the absence of other triads which would compensate for the distorting effects of this triad.

Method

Another experiment was warranted by the results of Experiment 1, which indicated that balanced incomplete block designs could provide reliable approximations to the configuration emerging from a complete design. This time it was desirable to look at a larger domain, one in fact large enough so that it would have been extremely cumbersome to administer all possible triads while keeping the sample constant. In short, it was time to test the efficacy of balanced incomplete block designs in the very situation in which they might be of some use.

Defining the Domain

Kinship, the domain chosen, is one that has been studied previously (Burton, 1968; Romney & D'Andrade, 1964; Wexler & Romney, 1969). *Father*, *son*, *brother*, *grandson*, *grandfather*, *uncle*, *nephew*, *cousin*--the eight male kin terms studied in these papers--were supplemented by the following female kin terms: *mother*, *daughter*, *sister*, *granddaughter*, *grandmother*, *aunt*, *niece*. The total number of items considered was 15.

Experiment 2

Subjects and Stimuli

Seventy-eight male and female undergraduate students at the University of California at Irvine were presented with 105 triads composed of three balanced incomplete block designs of 35 triads each for $\underline{n} = 15$, where $\lambda = 1$.

Analysis

The triads data were analyzed by Shepard-Kruskal multidimensional scaling methods using the TORSCA program (Kruskal, 1964a, 1964b; Shepard, 1964a, 1964b; Young and Torgerson, 1967). (One of the balanced designs was rerun using the Kruskal program as it appeared that the output configuration was a local minimum, which, in fact, it was.)

Generating the incomplete balanced block designs for $n = 15$, $\lambda = 1$.

Although a cyclic solution is possible for $\lambda = 3$ (seven cycles of 15 triads each), no such solution is possible for $\lambda = 1$. Consequently, an alternative procedure was used to find three non-overlapping designs for $\lambda = 1$ which were then combined to produce the design for $\lambda = 3$.[2] An outline of this procedure with an example follows.

The first step is to partition the 15 objects into five sets of three each. Label the sets A, B, C, D, and E. These constitute the first five triads of the design and form the base for generating the other 30 triads:

A. 1,2,3
B. 4,5,6
C. 7,8,9
D. 10,11,12
E. 13,14,15

Note that not only the identity of elements in each base set, but also the order of the elements within each set can determine the designs which are generated, i.e., if the set A were (3,2,1) rather than (1,2,3), the final outcome would be different. After the base sets have been delineated, each additional triad has to include members from each of three different base sets so that there will be no duplication of the pairings in any of the base sets. There are ten different ways to take three of these sets at a time:

---

[2] Schucker presents a fourth solution for $n = 15$, $\lambda = 1$ which overlaps with all three of our solutions. (See Appendix)

ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE.

If we can select three triads from each of these ten combinations of sets such that no pairings are duplicated with those of the base sets, we will have the remaining 30 triads.

Consider the pair of sets (A,B). Within these two sets, the following pairings of items are needed:

(1) 1 to 4, 2 to 5, and 3 to 6 (Space zero, i.e., where the first member of set A is paired with the first member of set B; the second member of set A is paired with the second member of set B; and the third member of set A is paired with the third member of set B.)

(2) 1 to 5, 2 to 6, and 3 to 4 (space one, i.e., where the first member of set A is paired with the second member of set B; the second member of set A is paired with the third member of set B; and the third member of set A is paired with the second member of set B.)

(3) 1 to 6, 2 to 4, and 3 to 5 (space two, i.e., where the first member of set A is paired with the third member of set B; the second member of set A is paired with the first member of set B; and the third member of set A is paired with the second member of set B.)

The pair of sets (A,B) occurs in three combinations: ABC, ABD, and ABE. It will be necessary for each one of these three combinations to take care of a different one of the three spacings so that all three spacings will be covered by the three combinations. For example, one possibility for the three triads for ABC is (1,5,7), (2,6,8), and (3,4,9). Simple trinary arithmetic shows that these triads take care of space one for the pair (A,B) and also cover space two for the pair (B,C) and space zero for the pair (A,C): one (A,B) + two (B,C) = zero (A,C).

The problem of constructing the design reduces to finding combinations of orderings which satisfy the criterial constraints so that each pair of cues in the entire design has been covered by spacings of zero, one, and two. It is helpful to keep track of which pairs have satisfied the various spacings with a table. After choosing the above triads for ABC, this table would look as follows:

	A	B	C	D	E
A		1	0		
B			2		
C					
D					
E					

When a balanced design is completed, the table should look like this:

	A	B	C	D	E
A	X	0,1,2	0,1,2	0,1,2	0,1,2
B		X	0,1,2	0,1,2	0,1,2
C			X	0,1,2	0,1,2
D				X	0,1,2
E					X

We know of no algorithm for achieving this happy state of affairs; it is simply a matter of trial and error. Very often, one will approach the end of generating a balanced design of triads only to find that the incomplete design cannot be completed. This situation indicates that a wrong branch was taken in the decision tree early in the stages of triad selection.

In order to obtain three completely non-overlapping designs, it was necessary to generate about 15 different balanced designs. The resulting designs are as follows:

Design 1.

Base Sets

A: 1,2,3

B: 6,5,4

C: 8,7,9

D: 12,11,10

E: 14,15,13

Derived Sets

ABC	(1,5,8)(2,4,7)(3,6,9)
ABD	(1,6,11)(2,5,10)(3,4,12)
ABE	(1,4,14)(2,6,15)(3,5,13)
ACD	(1,7,12)(2,9,11)(3,8,10)
ACE	(1,9,15)(2,8,13)(3,7,14)
ADE	(1,10,13)(2,12,14)(3,11,15)
BCD	(6,8,12)(5,7,11)(4,9,10)
BCE	(6,7,13)(5,9,14)(4,8,15)
BDE	(6,10,14)(5,12,15)(4,11,13)
CDE	(8,11,14)(7,10,15)(9,12,13)

Design 2.

Base Sets

A: 1,2,10

B: 6,4,15

C: 9,3,7

D: 12,8,11

E: 13,14,5

Derived Sets

ABC (1,4,9)(2,15,3)(10,6,7)

ABD (1,12,15)(2,6,8)(4,10,11)

ABE (1,6,14)(2,4,5)(10,13,15)

ACD (1,3,8)(2,7,11)(9,10,12)

ACE (1,7,13)(2,9,14)(3,5,10)

ADE (1,11,5)(2,12,13)(10,8,14)

BCD (3,6,12)(4,7,8)(9,11,15)

BCE (5,6,9)(3,4,13)(7,14,15)

BDE (6,11,13)(4,12,14)(5,8,15)

CDE (9,8,13)(3,11,14)(7,12,5)

Design 3.

Base Sets

A: 4,5,13

B: 7,9,3

C: 6,10,12

D: 11,14,15

E: 8,1,2

Derived Sets

ABC	(4,7,12)(5,6,3)(8,10,13)
ABD	(4,12,8)(5,9,11)(7,13,14)
ABE	(4,9,2)(5,7,[illegible])(13,1,3)
ACD	(4,6,11)(5,10,14)(12,13,15)
ACE	(4,10,1)(5,[illegible],2)(6,[illegible],13)
ADE	(4,8,14)(5,15,1)(11,13,2)
BCD	(6,9,15)(7,10,11)(12,14,3)
BCE	(8,9,12)(6,7,1)(10,2,3)
BDE	(9,14,1)(7,15,2)(8,11,3)
CDE	(12,11,1)(6,14,2)(8,10,15)

Results of Experiment 2

The kin term data were scaled in four, three, two and one dimensions. Table 9 shows the stress values of solutions in different dimensions for the three sets of λ = 1 data and for the λ = 3 data (which are a composite of these data).

Insert Table 9 about here

The stress patterns are virtually identical for the four cases. In Figure 9, stress is plotted against the number of dimensions for the λ = 3 case.

Insert Figure 9 about here

There is no clear point at which stress ceases to decrease; rather, with each dimension, there is a small decrease. However, for the English kin terms, the theory of componential analysis as set forth by Romney and D'Andrade (1964) predicts a four-dimensional structure for male and female English kin terms, where the dimensions should be (1) direct-collateral, (2) reciprocity, (3) generation, and (4) sex. Table 10 shows a list of the kin terms with their assigned numbers and their values on each of the four dimensions.

Insert Table 10 about here

Examination of the four-dimensional configuration for $\lambda = 3$ reveals that these dimensions are all present.

In the Torsca program, a latent root analysis is performed before the nonmetric multidimensional scaling. The final axes of the multidimensional configuration tend to correspond to the dimensions which are uncovered through this analysis: most of the discovery of structure occurs through the latent root analysis. By examining the size of the latent roots, or eigen values, one can get an idea of the relative importance of the different dimensions. For the kinship in four dimensions, $\lambda = 3$, the first nine eigen values are 2,369, 1.370, 1.300, 1.263, .662, .438, .269, .202, 0.000, indicating that the first four dimensions are much more important than the remaining four. The first of these four is considerably more important than the others, and the last three of the four are all equally important.

For these reasons, i.e., our theory that the "true" structure is four-dimensional combined with the presence of these same four dimensions in the scaling, and the high eigen values that these dimensions took in the latent root analysis, four-dimensional configurations were used to represent the kin terms. Figure 10 is a graph of dimension 1 against dimension 2 for $\lambda = 3$. Figure 11 is a graph of dimension 3 against dimension 4 for the same configuration.

Insert Figures 10 and 11 about here

Table 11 lists the rank correlation coefficients among the various configurations for $\lambda = 1$, 2, and 3.

Insert Table 10 about here

From this table it may be inferred that the results for $\lambda = 1$ are not very reliable, but that with $\lambda = 2$, they have become stabilized. This point is illustrated in Figures 12 through 15 which show graphs of dimensions 1 versus 2 (Figure 12) and of dimensions 3 versus 4 (Figure 13) of the "worst" configuration for $\lambda = 2$ (which has a rank correlation to $\lambda = 3$ of .895); and graphs of dimensions 1 versus 2 (Figure 14) and of dimensions 3 versus 4 (Figure 15) of the "best" configuration for $\lambda = 1$ (which has a rank correlation to $\lambda = 3$ of .716).

Insert Figures 12 through 15 about here

The $\lambda = 1$ configuration preserves the distinction between direct and collateral, but it shows increasing amounts of distortion in succeeding dimensions. Thus in the generation dimension, grandmother is drawn up towards the first generation kin terms; and son is pulled towards the second generation kin terms. In the sex dimension, grandfather has been included among the Females. In the reciprocity dimension, reciprocal pairs are lined up approximately along the north-south axis (except for grandfather-grandson), but son is more positive than father, whereas in the other cases the older relative is more positive.

A possible explanation for the misplacement of *grandfather* on dimension 3 is the presence of the triad *grandmother*, *father*, *grandfather*. This triad has had the effect of increasing the distance from *father* to *grandfather* so that in $\lambda = 3$ that distance is the fifty-sixth largest, but in this $\lambda = 1$ it is the fifth largest. The distance between *grandfather* and *grandmother* has remained unchanged; in the $\lambda = 3$ configuration, it is ninety-eighth largest and hundreth largest for $\lambda = 1$. Through the presence of this triad, *grandfather* lost one of the critical trials on which it w could be classified with males. Another triad suspected of contributing in the same manner to the misplacement of *grandfather* is *granddaughter*, *grandfather*, *brother*. For $\lambda = 3$ the distance *grandfather-brother* is twelfth largest, but in this $\lambda = 1$ design it is the largest distance in the structure. Correspondingly, for $\lambda = 3$, the distance *grandfather-granddaughter* is eighty-first, whereas for $\lambda = 1$, it is one hundred and first. So the triad *granddaughter*, *grandfather*, *brother* has reduced the distance from *grandfather* to *granddaughter* (a female) and increased the distance from *grandfather* to *brother* (a male).

The reversal in reciprocity for *son-father* may be explained by the triad *niece*, *son*, *uncle* which would lead people to classify *son* with *uncle* on the criterion of sex, thus reducing the distance from *son* to *uncle*, and also causing the reciprocity distinction to be overridden, which would result in *son* being placed with *niece*. The distance *son-uncle* is thirty-ninth largest in $\lambda = 3$, but seventieth largest in $\lambda = 1$. Another such triad is *nephew*, *father*, *son* which will also cause reciprocity to be overridden in favor of the direct-collateral dimension. The relative distances are

	$\lambda = 3$	$\lambda = 1$
nephew-son	25	42
father-son	95	95

The relative distance from father-son is unchanged, but the distance from son to nephew is increased for $\lambda = 1$. There could be evidence of this change on the reciprocity dimension.

The distortion in the generation dimension, i.e., grandmother is split off from the other second generation kin terms, could be explained by the triad grandson, granddaughter, grandmother which could increase the distance between grandmother and granddaughter. The relative distances are

	$\lambda = 3$	$\lambda = 1$
granddaughter-grandson	54.5	61.5
grandmother-granddaughter	85.0	61.5

In $\lambda = 1$, the two distances are identical, whereas in $\lambda = 3$ the distance grandmother-granddaughter is much smaller.

Similarly, the southward displacement of son in the generation dimension can be partially accounted for by the triad granddaughter, son, daughter which we can predict will increase the distance son-daughter. The relative distances are

	$\lambda = 3$	$\lambda = 1$
son-daughter	84	58.5
granddaughter-grandson	89	55

Both distances are greater in $\lambda = 1$. Thus the prediction about son-daughter is confirmed, but it is not simply the case that this triad is solely responsible.

The "best" configuration for $\lambda = 1$ discussed above was produced by design 1 (see Figures 14 and 15). The "worst" configuration for $\lambda = 2$ was formed by combining these data with those from design 2. In this configuration, unlike in the others, sex is the second rather than the third dimension. The TORSCA program does a latent root analysis between the nonmetric multidimensional scaling; this result indicates that the latent root for the sex dimension was larger than the one for the generation dimension. The only appreciable difference between this result and the result for $\lambda = 3$ is that the generation dimension is not quite as clear. However, most of the problems which existed with the $\lambda = 1$ scaling have been solved.

Discussion and Summary

The purpose of this paper has been to investigate ways to reduce experimental labor with the triads test for judged similarity. In order to complete this investigation it has been necessary to make hypotheses about the processes involved in the triads test. These hypotheses enter into our conclusions about the uses of balanced incomplete block designs to reduce the number of triads per test.

In both of our experiments, the designs for $\lambda = 2$ had a satisfactory degree of reliability, whereas the designs for $\lambda = 1$ did not. This judgment is based on two kinds of evidence. First, there was a substantial increase in rank correlations between the two cases. Secondly, the output configuration for $\lambda = 2$ "looked" substantially the same as the configuration for all triads (vegetables) or as the theoretical structure (kinship), whereas some (vegetables) or all (kinship) of the $\lambda = 1$ configurations had significant departures from those models. The two experiments provide empirical evidence that the triads test is reliable with balanced incomplete block designs for $\lambda = 2$.

It is also possible to make a theoretical argument that designs for $\lambda = 2$ will correct the distortions which occur with designs for $\lambda = 1$. For the cluster structure of the vegetables, distortions occurred when there was a triad which contained the three members of a cluster. For the dimensional structure of kinship, distortions could occur more easily because there were more possibilities for sets of three items to be proximate to each other within the structure. However, in this case, each such triad had a smaller distorting effect. In either case, three items which were proximate in the structure were contrasted only with themselves and with no items which were more different to any of them than they were to themselves. In either case, this contrast to external elements could not occur because with $\lambda = 1$, all triads in which these items could appear had been included.

Take the simple example of a cluster of three items. With $\lambda = 1$, suppose the three items comprise a triad. Then in the $\lambda = 1$ configuration, that cluster will blow up. However, for $\lambda = 2$ there must be another triad for each of the pairs of items in the cluster which contrasts that pair to some external element. Those three other triads will partially compensate for the distorting effects of the first triad, so that there will be much less distortion in the configuration for $\lambda = 2$. In the cases for which this happened with the vegetables, the distortion for $\lambda = 2$ was negligable.

Suppose that clusters had four members rather than three. For example suppose the four elements, A,B,C, and D had a true spatial relationship as follows:

A B

C D

In the λ = 1 design it is possible to have a triad such as A,B,C which will tend to scatter the cluster. The balanced design must include every pair once. This triad includes the pairs AB,BC, and AC. It must also have triads which include the pairs AD,BD, and CD. These must be contrasted to items which are external to the cluster, and will tie A,B, and C to a position proximate to D, thus helping to preserve the cluster. The result of this balanced design for λ = 1 might be as follows:

C B

D

A

The cluster is maintained intact, but its internal structure is lost. If this hypothetical design were added to another one to produce a λ = 2 design, the changes are that the new triads would all contrast pairs within the cluster to items outside the cluster, thus strengthening the identity of the cluster. However, it is possible that the new design would include another triad, such as ABD, which would also tend to distort the cluster. But, if that happened, there would also have to be another three triads which contrasted (in this example) AC,BC, and CD to elements outside the cluster. The net result would be that similarity between A and B would be very low, since A and B were included in the two distorting triads ABC and ABD. The CD similarity would be very high, since the pair CD was included in neither of those two triads. Other similarities within the cluster would have intermediate values since they would be included in one but not both of the two triads. Thus, the final cluster would be intact, but would be distorted into a trapezoid shape:

A B

C D

It should be emphasized here that the worst kind of distortion that can happen with the triads test has been eliminated by the selection of items for the test which all belong to the same semantic domain. This removes the possibility of any of the items residing simultaneously in two disjoint spaces. Thus, if in a triads test *light* were a member of a triad, *light*, *dark*, *dusk*; there would not also appear in the same test the triad *light*, *heavy*, *medium* because of the restriction that all elements be of the same domain. *Light* would not be a part of two contexts in each of which it was defined on different sets of features (Stefflre, 1969).

Fundamental to the possibility that balanced block designs can preserve a structure while reducing the redundancy of the triads test is the notion that the test is indeed redundant; that there are extra presentations of each pair of items. In order for this to be true, the test must be context-free; that is, the results from the different presentations of a pair must not depend upon the identity of the third item. We have seen that this is not completely true. The context problems which can occur by including items from separate domains can be solved by restricting the analysis to members of a single domain. The context problems which can occur when triads include members of a single cluster, or items which are close in the space; can not be handled so easily. However, we have included some speculations as to why these problems are much less likely to happen for $\lambda = 2$ than they are for $\lambda = 1$. We can also use what we know about the distorting effects of such triads to aid in the experimental process.

Suppose we have an *a priori* hypothesis about the structure of a semantic domain. We can use that hypothesis to aid in the selection of the particular block design. If the first design which is chosen includes triads all of whose members are predicted to be proximate in the structure;

we can reject it in favor of another design which does not have that problem. (From a given design we can generate a large number of others simply by permuting the assignment of numbers to items.) If the hypothesis is true, this design will preserve the structure better than the original one. If it is not true, we have done no more damage to the structure than with any other design chosen at random.

In many cases, however, we have no a priori hypothesis, but want to use the triads test as a procedure to discover the structure. This would usually be the case in anthropological field work. In this case, it is still possible to use what is known about the ways in which particular triads can distort a structure as a guide to interpretation of the scaled configuration from triads data. In this case, however, it would be better to have data from at least two different balanced block designs. It may then be possible to explain the ways in which they differ from each other in terms of the presence of triads whose members are all proximate in the true structure, and thus to deduce what that structure is. That is, certain triads can be presumed to have caused transformations upon the structure, and it is possible to reverse the transformations to deduce the nature of the structure. That this is possible is demonstrated by the analysis of reasons why the $\lambda = 1$ configuration for kinship deviated from the $\lambda = 3$ configuration. In all cases, it was possible to predict the effect of a single triad on the rank ordering of distances in the $\lambda = 1$ configuration as compared with the rank ordering of distances in the $\lambda = 3$ configuration. Thus, given the presence of a particular triad, one can deduce its effect on the configuration and work backwards towards a plausible hypothesis about the true structure. It might then be desirable to design another experiment as a retest of the hypothesis. For a large number of items, several triads tests with a balanced design for $\lambda = 2$ would still have a smaller number of triads than a single complete test; and would have the advantage of providing several different analyses of the same domain. It is possible to pool them and analyze them separately, and obtain a clear idea as to which distortions have occurred and where. In this manner. the use of balanced block designs can lead to a reliable model of the structure of a semantic domain, while reducing experimental labor. It thus makes possible the use of the triads test in large semantic domains than was previously possible.

APPENDIX

Below is a diagram showing the pattern of solutions for incomplete balanced block designs to be applied to triads data for ns between 9 and 25. The shaded cells indicate for a given n those λs for which designs are not possible. The cells which have numbers in them indicate solutions available to the authors either having been generated by them or published. The numbers in those cells are the number of triads for that solution. The blank cells are those for which designs are theoretically possible but for which solutions have not been generated, to our knowledge. Some solutions, e.g., n = 20, λ = 6, both among those that are available and those that are not are impractical because of the large number of triads generated by them.

	λ					
n	1	2	3	4	5	6
7	7	14	21	28	35	42
8	▨	▨	▨	▨	▨	▨
9	12	24	36	48	60	72
10	▨	30	▨	60	▨	
11	▨	▨	55	▨	▨	
12	▨		▨		▨	
13	26	52	78			
14	▨	▨	▨	▨	▨	
15	35	70	105			
16	▨	80	▨		▨	
17	▨	▨	136	▨	▨	
18	▨		▨		▨	
19	57	114				
20	▨	▨	▨	▨	▨	380
21	70		210			
22	▨		▨		▨	
23	▨	▨	253	▨	▨	
24	▨		▨		▨	
25	100	200				

All the solutions for 9 and some of them for 15 have been presented in the body of the paper. The following is a presentation of solutions for n = 10 through n = 15 with additional alternative solutions for some λs for some ns.

1. n = 10.

a. λ = 2, 30 triads (solution from Cochrane and Cox, 1957, p. 540)

	Rep.				Rep.				Rep.		
Set	I	II	III	Set	IV	V	VI	Set	VII	VIII	IX
1)	1	2	3	11)	1	2	4	21)	1	3	5
2)	2	5	8	12)	2	3	6	22)	2	7	6
3)	3	7	4	13)	3	4	8	23)	3	8	9
4)	4	1	6	14)	4	9	5	24)	4	2	10
5)	5	8	7	15)	5	7	1	25)	5	6	3
6)	6	4	9	16)	6	8	9	26)	6	1	8
7)	7	9	1	17)	7	10	3	27)	7	9	2
8)	8	10	2	18)	8	1	10	28)	8	4	7
9)	9	3	10	19)	9	5	2	29)	9	10	1
10)	10	6	5	20)	10	6	7	30)	10	5	4

b. λ = 4, 60 triads

cycle 1: 1,4,5 step 1 to 10,3,4

cycle 2: 1,5,6 step 1 to 10,4,5

cycle 3: 1,3,5 step 1 to 10,2,4

cycle 4: 1,3,6 step 1 to 10,2,5

cycle 5: 1,4,7 step 1 to 10,3,6

cycle 6: 1,2,3 step 1 to 10,1,2

2. n = 11

λ = 3, 55 triads

cycle 1: 1,3,4 step 1 to 11,2,3

cycle 2: 1,5,6 step 1 to 11,4,5

cycle 3: 1,3,7 step 1 to 11,2,6

cycle 4: 1,4,5 step 1 to 11,3,4

cycle 5: 1,3,6 step 1 to 11,2,5

3. n = 13

a. λ = 1, 26 triads (both from Cochran and Cox, 1957, pp. 477-478).

(1) cycle 1: 1,4,5 step 1 to 13,3,4

cycle 2: 1,6,8 step 1 to 13,5,7

(2) cycle 1: 1,2,5 step 1 to 13,1,4

cycle 2: 1,3,8 step 1 to 13,2,7

b. λ = 2, 52 triads

Combine 3.a.(1) with 3.a.(2).

4. n = 15 (also see body of paper).

a. λ = 1, 35 triads

(1) Base Sets

1,2,3

4,5,6

7,8,9

10,11,12

13,14,15

Derived Sets

1,5,8
1,6,11
1,4,14
1,7,12
1,9,15
1,10,13
6,8,12
6,7,13
6,10,14
8,11,14

2,4,7
2,5,10
2,6,15
2,9,11
2,8,13
2,12,14
5,7,11
5,9,14
5,12,15
7,10,15

3,6,9
3,4,12
3,5,13
3,8,10
3,7,14
3,11,15
4,9,10
4,8,15
4,11,13
9,12,13

(2)

Base Set

1,2,10
6,4,15
9,3,7
12,8,11
13,14,5

Derived Sets

1,4,9
1,12,15
1,6,14
1,3,8
1,7,13
1,11,5
3,6,12
5,6,9
6,11,13
9,8,13

2,15,3
2,6,8
2,4,5
2,7,11
2,9,14
2,12,13
4,7,8
3,4,13
4,12,14
3,11,14

10,6,7
4,10,11
10,13,15
9,10,12
3,5,10
10,8,14
9,11,15
7,14,15
5,8,15
7,12,5

(3) (Solution from Schucker, 1959, p. 275)

Set	Rep. I			Set	Rep. II			Set	Rep. III		
(1)	1	15	8	(6)	3	15	10	(11)	5	15	12
(2)	4	5	7	(7)	6	7	9	(12)	8	9	11
(3)	13	14	9	(8)	1	2	11	(13)	3	4	13
(4)	3	6	11	(9)	5	8	13	(14)	7	10	1
(5)	10	12	2	(10)	12	14	4	(15)	14	2	6

Set	Rep. IV			Set	Rep. V			Set	Rep. VI		
(16)	7	15	14	(21)	9	15	2	(26)	11	15	4
(17)	10	11	13	(22)	12	13	1	(27)	14	1	3
(18)	9	12	3	(23)	7	8	3	(28)	9	10	5
(19)	5	6	1	(24)	11	14	5	(29)	13	2	7
(20)	2	4	8	(25)	4	6	10	(30)	6	8	12

Set	Rep. VII		
(31)	13	15	6
(32)	2	3	5
(33)	11	12	7
(34)	1	4	9
(35)	8	10	14

b. $\lambda = 2$, 70 triads

Combine 4.a.(1) and 4.a.(2).

c. $\lambda = 3$, 105 triads

cycle 1: 1,3,7 step 1 to 15,2,6

cycle 2: 1,4,8 step 1 to 15,3,7

cycle 3: 1,2,3 step 1 to 15,1,2

cycle 4: 1,5,10 step 1 to 15,4,9

cycle 5: 1,4,9 step 1 to 15,3,8

cycle 6: 1,3,8 step 1 to 15,2,7

cycle 7: 1,2,8 step 1 to 15,1,7

5. $n = 16$

$\lambda = 2$, 80 triads

cycle 1: 1,4,9 step 1 to 16,3,8

cycle 2: 1,3,7 step 1 to 16,2,6

cycle 3: 1,2,8 step 1 to 16,1,7

cycle 4: 1,2,4 step 1 to 16,1,3

cycle 5: 1,5,10 step 1 to 16,4,9

6. $n = 17$

$\lambda = 3$, 136 triads

cycle 1: 1,5,10 step 1 to 17,4,9

cycle 2: 1,2,8 step 1 to 17,1,7

cycle 3: 1,3,6 step 1 to 17,2,5

cycle 4: 1,4,10 step 1 to 17,3,9

cycle 5: 1,5,11 step 1 to 17,4,10

cycle 6: 1,2,4 step 1 to 17,1,3

cycle 7: 1,3,10 step 1 to 17,2,9

cycle 8: 1,2,6 step 1 to 17,1,5

7. $\underline{n} = 19$

a. $\lambda = 1$, 57 triads

(1) cycle 1: 1,3,6 step 1 to 19,2,5

cycle 2: 1,8,9 step 1 to 19,7,8

cycle 3: 1,5,11 step 1 to 19,4,10

(2) cycle 1: 1,2,6 step 1 to 19,1,5

cycle 2: 1,4,11 step 1 to 19,3,10

cycle 3: 1,3,9 step 1 to 19,2,8

b. $\lambda = 2$, 114 triads

Combine 7.a.(1) and 7.2.(2).

(3) (From Cochran and Cox, 1957, p. 542)

cycle 1: 1,7,11 step 1 to 19,6,10

cycle 2: 1,2,13 step 1 to 19,1,12

cycle 3: 1,3,6 step 1 to 19,2,5

8. $\underline{n} = 20$

$\lambda = 6$, 380 triads

cycle 1: 1,4,11 step 1 to 20,3,10

cycle 2: 1,3,11 step 1 to 20,2,10

cycle 3: 1,5,11 step 1 to 20,4,10

cycle 4: 1,2,10 step 1 to 20,1,9

cycle 5: 1,5,10 step 1 to 20,4,9

cycle 6: 1,2,4 step 1 to 20,1,3

cycle 7: 1,3,8 step 1 to 20,2,7

cycle 8: 1,4,10 step 1 to 20,3,9

cycle 9: 1,8,16 step 1 to 20,7,15

cycle 10: 1,2,6 step 1 to 20,6,5

cycle 11: 1,3,9 step 1 to 20,2,8

cycle 12: 1,3,10 step 1 to 20,2,9

cycle 13: 1,2,7 step 1 to 20,1,6

cycle 14: 1,7,10 step 1 to 20,6,9

cycle 15: 1,5,12 step 1 to 20,4,11

cycle 16: 1,4,9 step 1 to 20,3,8

cycle 17: 1,2,5 step 1 to 20,1,4

cycle 18: 1,2,9 step 1 to 20,1,8

cycle 19: 1,3,7 step 1 to 20,2,6

9. n = 21

a. λ = 1, 70 triads (from Cochran and Cox, 1957, p. 479)

Set	Rep. I			Set	Rep. II			Set	Rep. III		
(1)	1	2	3	(8)	1	4	15	(15)	1	5	17
(2)	4	5	6	(9)	2	5	11	(16)	2	4	14
(3)	7	8	9	(10)	3	9	16	(17)	3	7	11
(4)	10	11	12	(11)	6	17	20	(18)	6	10	19
(5)	13	14	15	(12)	7	12	19	(19)	8	16	20
(6)	16	17	18	(13)	8	13	18	(20)	9	15	18
(7)	19	20	21	(14)	10	14	21	(21)	12	13	21

Rep. IV

Set			
(22)	1	6	9
(23)	2	7	16
(24)	3	8	21
(25)	4	17	19
(26)	5	10	13
(27)	11	15	20
(28)	12	14	18

Rep. V

Set			
(29)	1	7	21
(30)	2	13	17
(31)	3	10	18
(32)	4	8	11
(33)	5	16	19
(34)	6	12	15
(35)	9	14	20

Rep. VI

Set			
(36)	1	8	10
(37)	2	18	19
(38)	3	15	17
(39)	4	12	16
(40)	5	9	21
(41)	6	11	14
(42)	7	13	20

Rep. VII

Set			
(43)	1	11	18
(44)	2	10	20
(45)	3	5	12
(46)	4	9	13
(47)	6	16	21
(48)	7	14	17
(49)	8	15	19

Rep. VIII

Set			
(50)	1	12	20
(51)	2	6	8
(52)	3	14	19
(53)	4	18	21
(54)	5	7	15
(55)	9	10	17
(56)	11	13	16

Rep. IX

Set			
(57)	1	13	19
(58)	2	9	12
(59)	3	4	20
(60)	5	8	14
(61)	6	7	18
(62)	10	15	16
(63)	11	17	21

Rep. X

Set			
(64)	1	14	16
(65)	2	15	21
(66)	3	6	13
(67)	4	7	10
(68)	5	18	20
(59)	8	12	7
(70)	9	11	19

b. λ = 3, 210 triads

cycle 1: 1,4,10 step 1 to 21,3,9

cycle 2: 1,3,8 step 1 to 21,2,7

cycle 3: 1,5,13 step 1 to 21,4,12

cycle 4: 1,2,4 step 1 to 21,1,3

cycle 5: 1,6,12 step 1 to 21,5,11

cycle 6: 1,5,12 step 1 to 21,4,11

cycle 7: 1,2,10 step 1 to 21,1,9

cycle 8: 1,5,11 step 1 to 21,4,10

cycle 9: 1,2,9 step 1 to 21,1,8

cycle 10: 1,3,6 step 1 to 21,2,5

10. n = 23

λ = 3, 253 triads

cycle 1: 1,6,14 step 1 to 23,5,13

cycle 2: 1,5,12 step 1 to 23,4,11

cycle 3: 1,2,11 step 1 to 23,1,10

cycle 4: 1,4,10 step 1 to 23,3,9

cycle 5: 1,3,9 step 1 to 23,2,8

cycle 6: 1,8,12 step 1 to 23,7,11

cycle 7: 1,3,6 step 1 to 23,2,5

cycle 8: 1,7,10 step 1 to 23,6,9

cycle 9: 1,2,9 step 1 to 23,1,7

cycle 10: 1,2,6 step 1 to 23,1,5

cycle 11: 1,3,13 step 1 to 23,2,12

11. n = 25

a. λ = 1, 100 triads

(1) cycle 1: 1,2,13 step 1 to 25,1,12

cycle 2: 1,4,11 step 1 to 25,3,10

cycle 3: 1,5,10 step 1 to 25,4,9

cycle 4: 1,3,9 step 1 to 25,2,8

(2) cycle 1: 1,12,13 step 1 to 25,11,12

cycle 2: 1,8,11 step 1 to 25,7,10

cycle 3: 1,6,10 step 1 to 25,5,9

cycle 4: 1,7,9 step 1 to 25,6,8

b. λ = 2,200 triads

Combine 11.a.(1) with 11.a.(2).

BIBLIOGRAPHY

Burton, Michael L. *Multidimensional scaling of role terms* (Doctoral dissertation, Stanford University). Ann Arbor, Mich.: University Microfilms, 1968, No. 68-8160.

Burton, M. Semantic dimensions of occupation names. Paper presented at the Conference on Multidimensional Scaling in the Behavioral Sciences, Newport Beach, California, June 1969.

Cochran, W. G., and Cox, G. M. *Experimental designs.* New York: Wiley, 1957 (2nd ed.).

Coombs, C. H. *A theory of data.* New York: Wiley, 1964.

Durbin, J. Incomplete blocks in ranking experiments. *British Journal of Psychology, Statistical Section*, 1951, *4*, 85-90.

Gulliksen, H., and Tucker, L. R. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*, 1961, *26*, 173-83.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241-253.

Kelly, G. A. The psychology of personal constructs. New York: W. W. Norton, 1955.

Romney, A. K., and D'Andrade, R. G. Cognitive aspects of English kin terms. In A. K. Romney and R. G. D'Andrade (Eds.), *Transcultural studies in cognition. American Anthropologist Special Issue*, 1964 (*66* Pt.2, No. 3), 146-170.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1-27 (a).

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964, *29*, 115-129 (b).

Schucker, R. E. A note on the use of triads for paired comparisons. *Psychometrika*, 1959, *24*, 273-276.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962, *27*, 125-40 (a).

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, *27*, 219-246 (b).

Stefflre, V. J. Some applications of multidimensional scaling to social science problems. Paper presented at the Conference on Multidimensional Scaling in the Behavioral Sciences, Newport Beach, California, June 1969.

Torgerson, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

Wexler, K., and Romney, A. K. Individual variations in cognitive structures. Paper presented at the Conference on Multidimensional Scaling in the Behavioral Sciences, Newport Beach, California, June 1969.

Yates, F. A new method of arranging variety trials involving a large number of varieties. *Jour. Agr. Sci.*, 1936, *26*, 424-455.

Young, F. W., and Torgerson, W. S. TORSCA, a FORTRAN IV program for Shepard-Kruskal multidimensional scaling analysis. *Behavioral Science*, 1968, *12*, 498.

Table 1

Degree of Overlap Among Designs, λ = 1

	A	B	C	D	E	F	G
A	X						
B	2	X					
C	0	6	X				
D	2	5	0	X			
E	3	5	6	0	X		
F	1	3	1	1	3	X	
G	3	1	2	1	0	0	X

Table 1

[illegible] of Overlap [illegible] Designs, k = 2

	N	[illegible]	F	H	L
E*	E				
S	O*	R			
[illegible]	[illegible]	[illegible]	D		
K	U	[illegible]	S	F	
[illegible]	[illegible]	[illegible]	S	A	X

Table 3

Spearman Rank Correlation Coefficients
for Balanced Subsets where λ = 1

	ALL	A	B	C	D	E	F	G
A	.827	X						
B	.832	.822	X					
C	.952	.799	.909	X				
D	.742	.543	.614	.722	X			
E	.883	.678	.849	.909	.703	X		
F	.570	.390	.486	.565	.332	.602	X	
G	.716	.775	.707	.786	.454	.629	.380	X

Table 4

Spearman Rank Correlation Coefficients
for Balanced Subsets where $\lambda = 2$

	ALL	H	I	J	K	L
H	.873	X				
I	.898	.888	X			
J	.889	.796	.837	X		
K	.865	.807	.797	.912	X	
L	.764	.890	.735	.707	.693	X

Table 5

Spearman Rank Correlation Coefficients
for Balanced Subsets where $\lambda = 3$

	ALL	M	N	P
M	.954	X		
N	.945	.865	X	
P	.963	.949	.881	X

Table 6

Spearman Rank Correlation Coefficients
for Balanced Subsets where λ = 4

	ALL	Q	R	S
Q	.921	X		
R	.988	.892	X	
S	.961	.939	.937	X

Table 7

Spearman Rank Correlation Coefficients
for Random Sets

	All triads
12 triads	.402
24 triads	.590
36 triads	.808

Table 8

Spearman Rank Correlation Coefficients
for Two Sets of Split-Halves of Subjects

	ALL	ODDS	EVENS	FIRST	LAST
ODDS	.949	X			
EVENS	.978	.917	X		
FIRST 50	.904	X	X	X	
SECOND 50	.968	X	X	.904	X

Table [illegible]

Stresses for [illegible]

DIMENSIONS	[illegible]			
4D	.080	[illegible]	[illegible]	.090
3D	.145	[illegible]	[illegible]	.123
2D	.222	[illegible]	[illegible]	.216
1D	.396	[illegible]	[illegible]	.408
	λ = 3	[illegible]	[illegible]	λ = 1,
	Designs 1, 2, & 3	[illegible]	[illegible]	Design 3

Table 10

Values of Kin Terms on Four Dimensions

	Dimensions			
Kin Terms	DIRECT/ COLLATERAL	RECIPROCITY	GENERATION REMOVED FROM EGO	SEX
1. father	direct	father ↔ son	1	M
2. mother	direct	mother ↔ daughter	1	F
3. son	direct	son ↔ father	1	M
4. daughter	direct	daughter ↔ mother	1	F
5. grandfather	direct	grdfa ↔ grdso	2	M
6. grandmother	direct	grdmo ↔ grdda	2	F
7. nephew	collateral	nephew ↔ uncle	1	M
8. niece	collateral	niece ↔ aunt	1	F
9. cousin	collateral	none	0	M,F
10. uncle	collateral	uncle ↔ nephew	1	M
11. aunt	collateral	aunt ↔ niece	1	F
12. grandson	direct	grdso ↔ grdfa	2	M
13. granddaughter	direct	grdda ↔ grdmo	2	F
14. brother	direct	none	0	M
15. sister	direct	none	0	F

Table 11

Spearman Rank Correlations in Four Dimensions
for Kinship Configurations

	First	Second	Third	12	13	23	$\lambda = 3$
First	X						
Second	.372	X					
Third	.540	.393	X				
12	.735	.739	.552	X			
13	.782	.503	.776	.844	X		
23	.543	.727	.698	.792	.826	X	
$\lambda = 3$	.716	.625	.716	.895	.937	.903	X

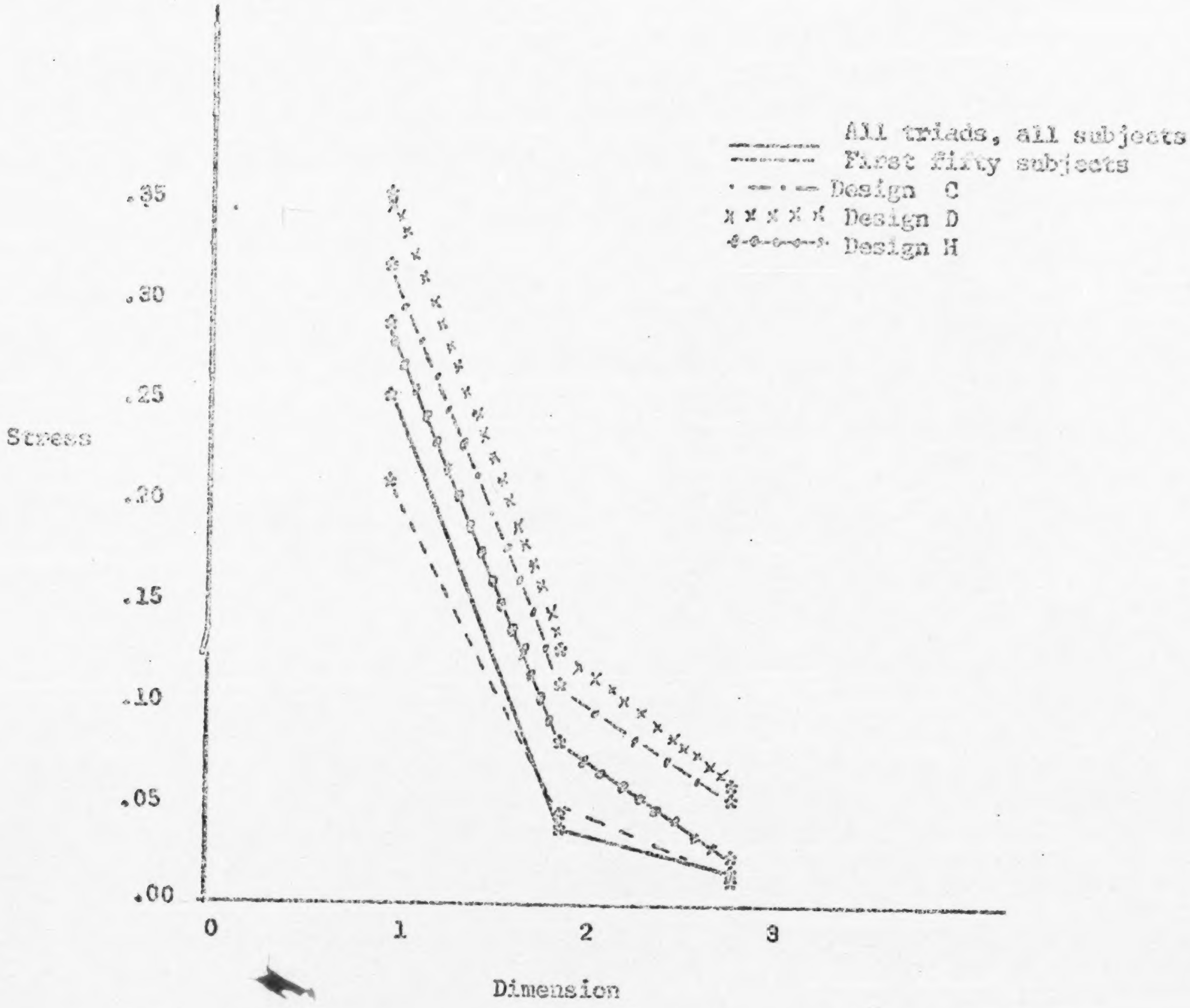


Fig. 1 . Goodness of fit of the multidimensional configurations of vegetables for complete triads for all subjects, and for the first 50 subjects; and for balanced designs C ($\lambda = 1$), D ($\lambda = 1$), and H ($\lambda = 2$).

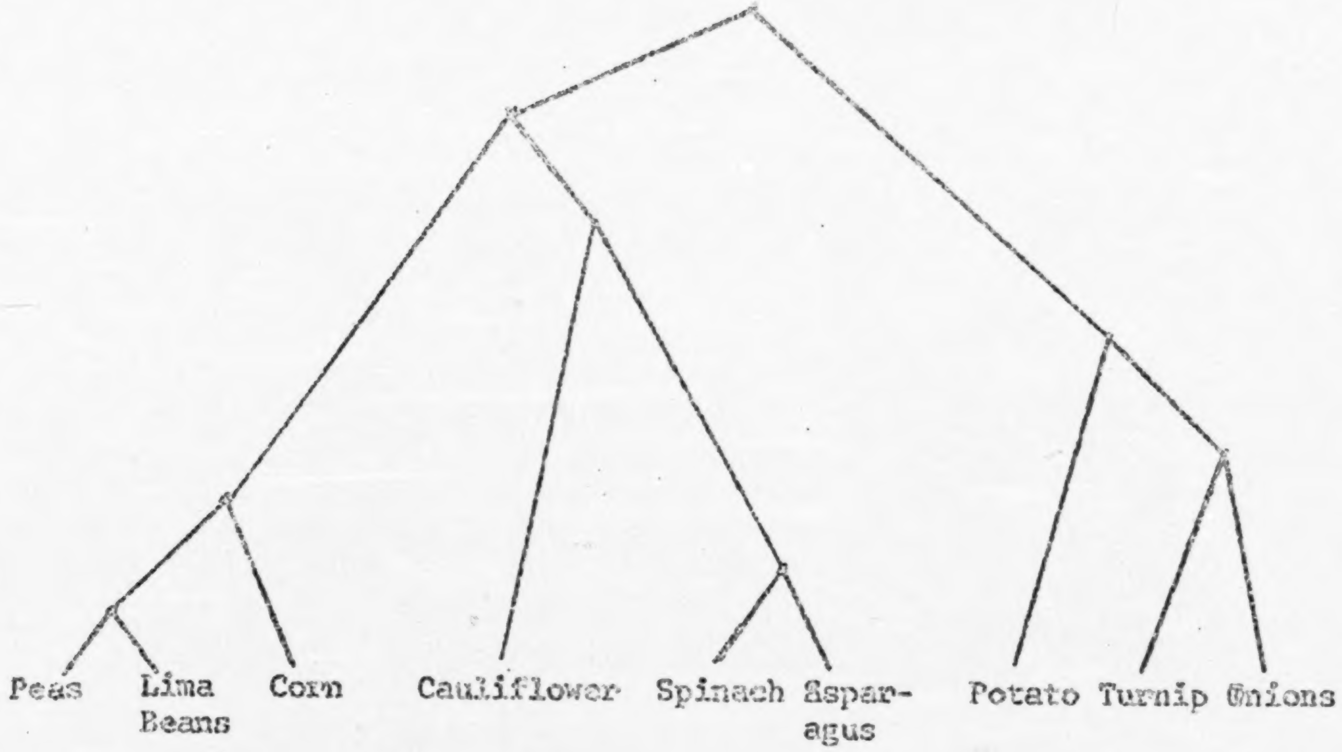


Fig. 2. Hierarchical clustering of vegetables, calculated by Johnson's method.

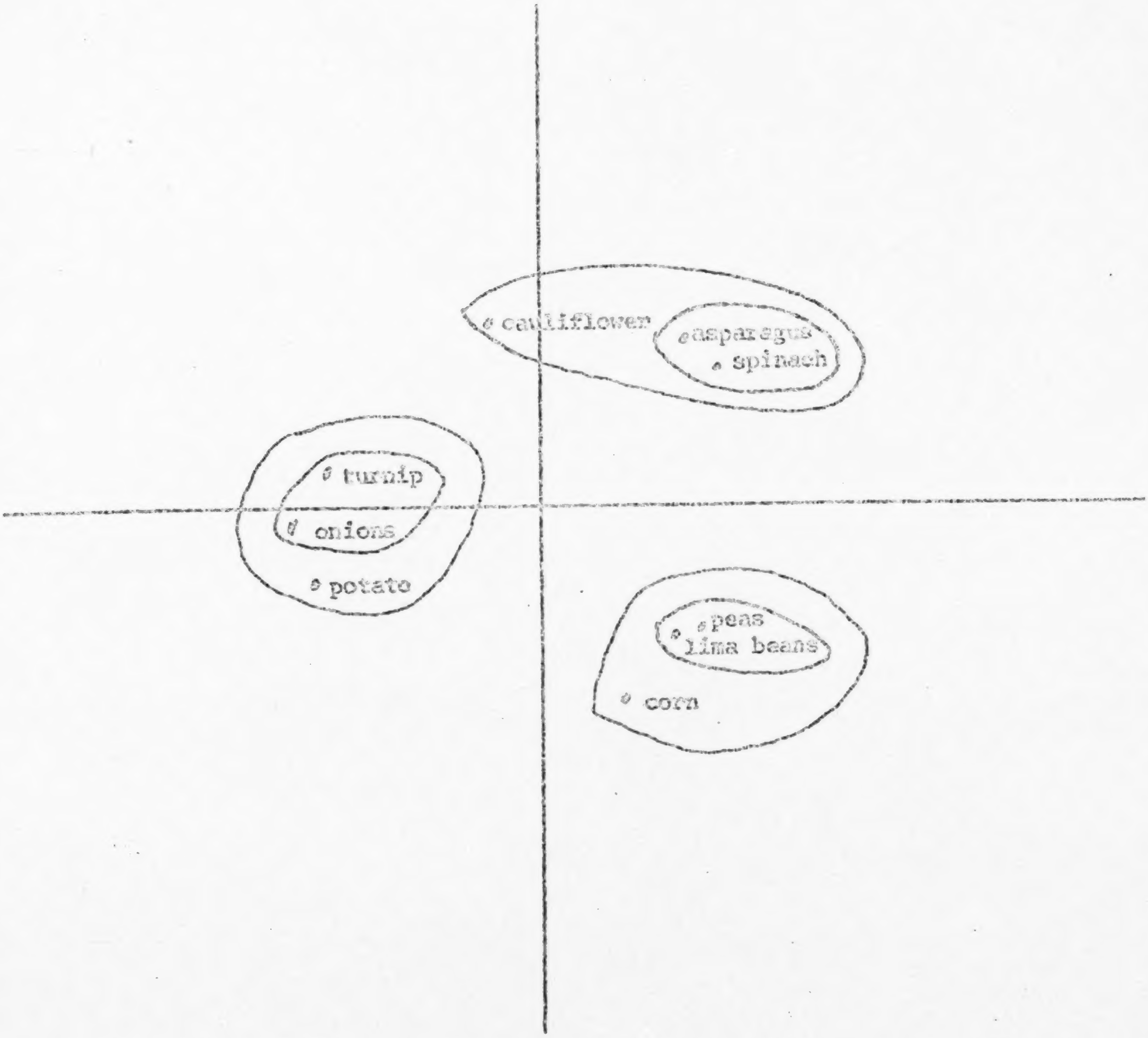


Fig. 3. Two-dimensional representation of nine vegetables.
All subjects, complete triads.

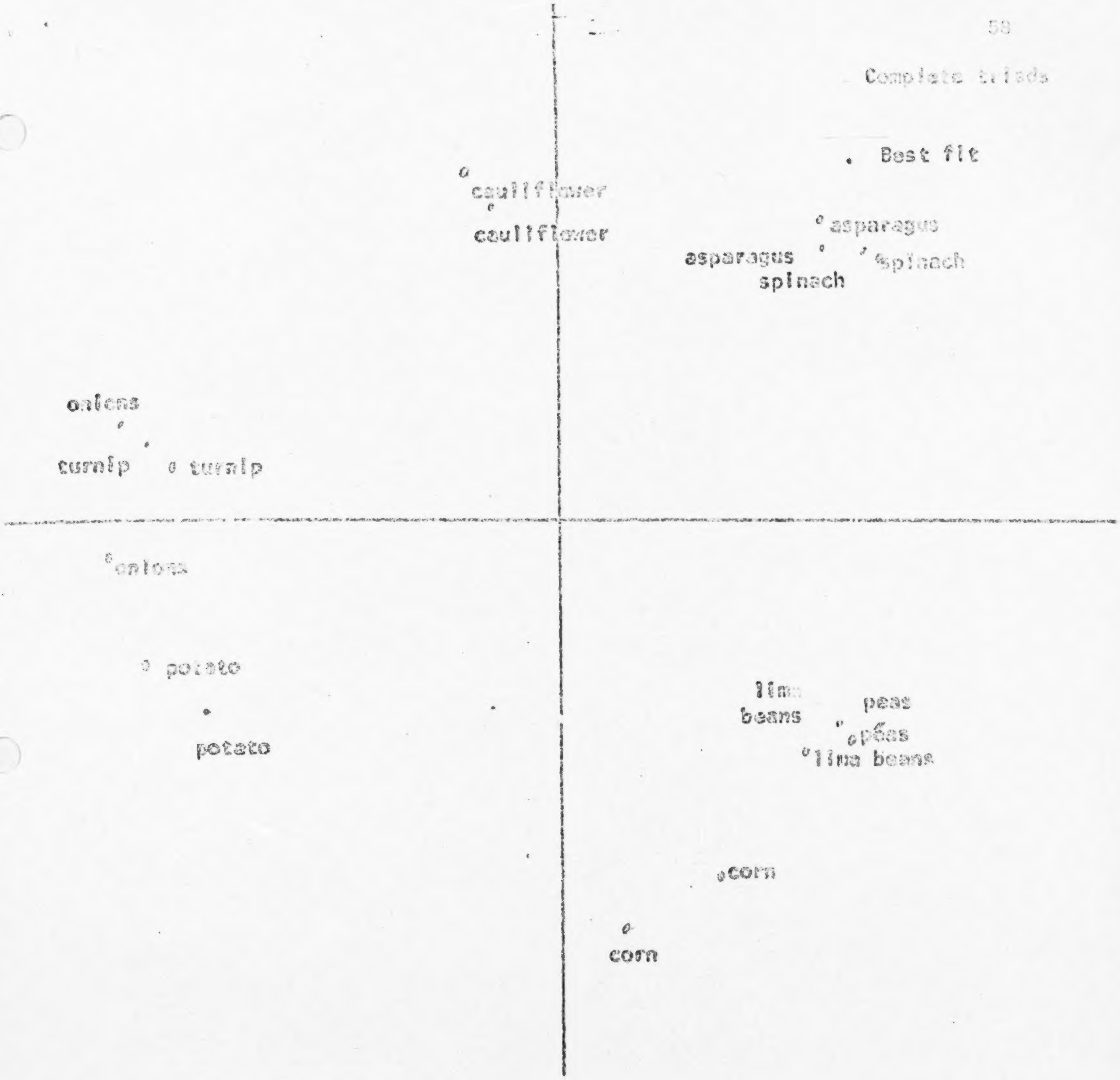


Fig. 4. Multidimensional scaling plot of design C the best fit for λ = 1, and design F, the worst fit for λ = 1, along with the complete triads configuration.

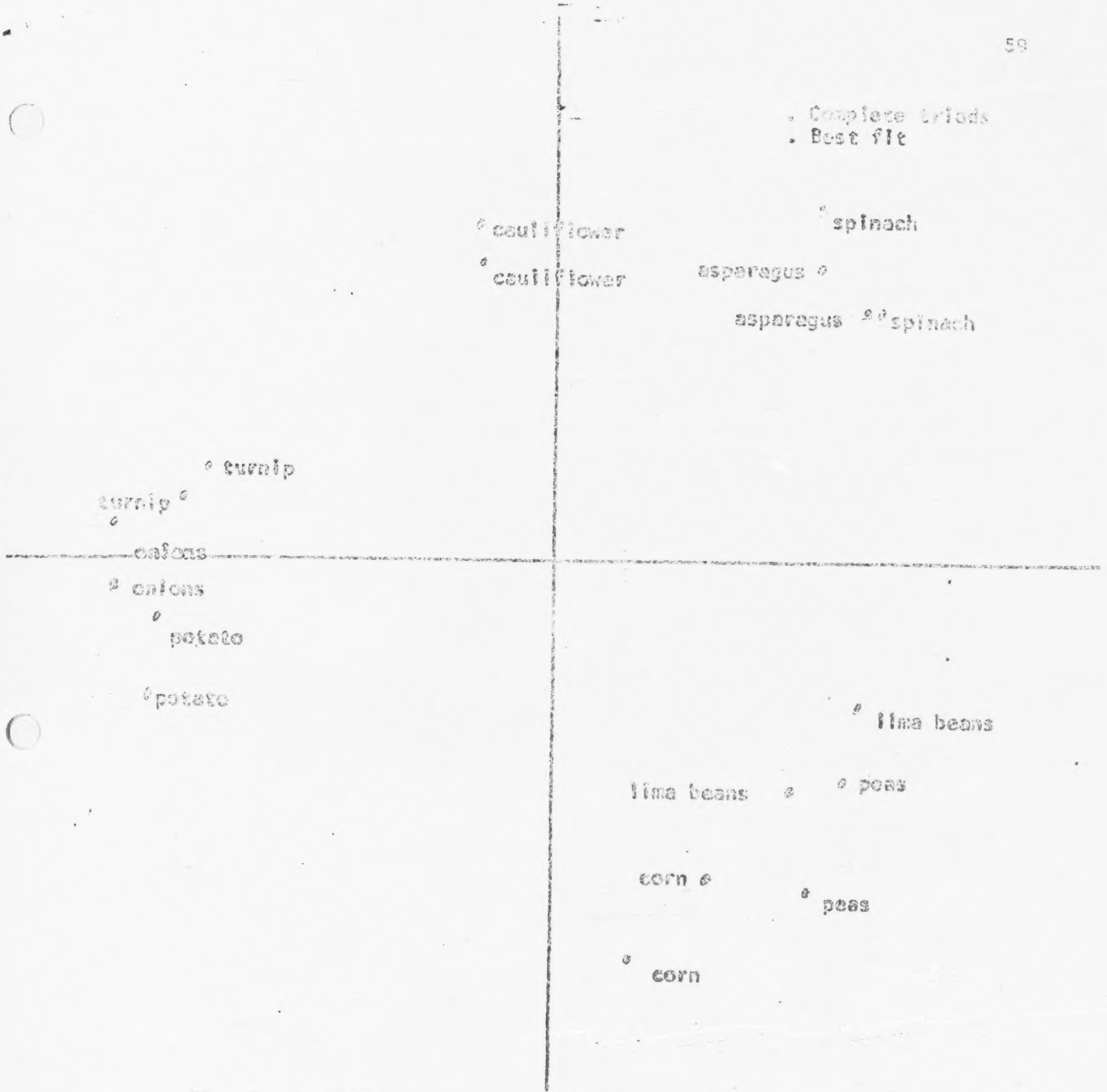


Fig. 5. Multidimensional scaling plot of design I, the best fit for $\lambda = 2$, and design L, the worst fit for $\lambda = 2$, along with the complete triads configuration.

How we understand negation[1]

Herbert H. Clark

Stanford University

Paper presented at Workshop on Cognitive Organization and Psychological Processes, University of California, Irvine, August 1970.

How we understand negation[1]

Herbert H. Clark

Stanford University

Negation is one of the most basic conceptual devices of language. It is so fundamental, in fact, that it is difficult to imagine either a natural or an artificial language that could exist without it. The uses of negation are numerous, the most obvious of which is in denials. Very often we know only enough about an event to be able to say what it is not; for example, we say *John isn't at home* because we cannot say for certain where John is--we only know where he is not. But another less obvious use of negation is in definitions of words. For example, *forget* means *not remember*; *out* means *not in*; *unhappy* means approximately *not happy*; and so on. If we broaden the definition of negation, however, negation can be found in many facets of human cognition. Indeed, the main aim of the present paper is to convince the reader that the study of negation and its comprehension has led to some fundamental insights not only into how people understand language, but also into how people think. More concretely, what I hope to do is demonstrate that negation can be treated in a unitary fashion despite its diverse guises in language and cognition.

Before continuing, I must list some of the different types of negation that we will be studying. The most obvious type, of course, is negation specified by a negative particle, as in *John isn't at home*, *Not everyone likes to eat courgettes*, *Rodhri has never seen Hadrian's Wall*, and so on. But English contains other subtler types of negation, as in *Hardly anybody eats rutabagas anymore*, *Few people here understand Sioux*, *Melanie forgot her suspenders*, and so on. The negation here is found in *hardly* (an "adverb"), *few* (a quantifier), and forgot (a verb). Even less obvious are the negatives implicit in absent, from, out, and the like, and, if it can even be called negation, the semantic markedness of such adjectives as small, short, shallow, and sad and of such prepositions as below, behind, and down. The first problem, then, is to find one conceptualization of negation that will fit all these cases, and the second is to construct a theory of comprehension that will account for all of them at the same time.

The number of psychological studies of negation is remarkably large. Nevertheless, their results have never been as clear cut as one would like. Beginning with the pioneering studies by Wason (1959, 1961, 1965; Wason & Jones, 1963), many psychologists have come up with one principal finding: negation is more difficult to comprehend than affirmation. But beyond this finding, the studies (to be reviewed later) have varied widely in their results. Some verification studies, for example, have found that negative sentences that are true can be verified faster than false ones, whereas other studies have found just the reverse. Also, some studies have shown negatives to be very much harder than affirmatives to comprehend, whereas others have found negatives to be only slightly harder than affirmatives. So the immediate problem arises as to how to reconcile all these and other conflicting results on the comprehension of negation.

The thesis of the present paper is just this: The comprehension of all forms of negation can be accounted for by a single unified model of comprehension--a model I will call the "true" model of negation. ("True" here should not be construed as a claim about the ultimate correctness of the model; rather, the "true" model will be later contrasted with a "conversion" model of negation, which in truth is a dishonest way of comprehending negation.)

I will show that the "true" model, when fully specified, accounts for most of the conflicts in the previous results and can be generalized to account for other types of linguistic and perceptual comprehension as well. The "true" model itself, however, is built up on two more basic concepts: (1) the semantic notion of presupposition; and (2) the processing notion of congruence. So the plan of the present paper is as follows. First, I will try to characterize the various types of linguistic negation and the role presupposition and other semantic notions play in these types. Second, I will present the "true" method of negation, showing how it relies heavily on a principle of congruence. Third, I will then review the previous ttudies of negation, including several studies by myself, a colleague, and several students that are not yet published. And finally, I will summarize several hypotheses about comprehension--and cognition more generally--that are suggested by the study of negation.

Types of negation

The sentence *Burton isn't at home* denies that Burton is at home. When would it be appropriate to make such a denial? If I thought that you had expected Burton to be home, or had said so, or had implied so in what you had just said, I would be very likely to say *Burton isn't at home*. Note that it seems appropriate to say *I know you think Burton is at home, but he isn't at home*, but quite inappropriate to say *I know you think Burton is at school, but Burton isn't at home*. These and many other linguistic examples show that a speaker makes an assumption about the beliefs (or apparent beliefs) of his listener whenever he utters a denial. Specifically, he assumes that the listener does or could well believe in the truth of what is being denied. In saying Burton isn't at home, the speaker assumes for some reason that the listener believes that Burton is or might be at home. As Wason (1965) pointed out, The whale isn't a fish is a plausible negative since it is reasonable to assume that a listener might think that whales are fish; but The whale isn't a bird, though true, sounds highly incongruous, for who could ever believe that a whale is a bird.

All this discussion is pertinent to what people come to know from listening to a negative sentence. This immediate knowledge apparently falls roughly into two parts. First, the listener-comprehender knows what the speaker thinks the listener's beliefs are. That is, the listener knows what will be called the presupposition of the sentence. Second, the listener knows what the speaker is asserting to be true. He knows what will be called the assertion of the sentence. In illustration, Burton isn't at home expresses a presupposition and an assertion. The presupposition is that the speaker supposes that the listener believes that Burton is or might be at home, and the assertion is that that belief is false. We might present what a person has understood of sentence (1a), then, approximately as in (1b), the paraphrase of which is given in (1c):

(1) a. Burton isn't at home.

b. (false (suppose (Burton at home)))

c. It is false (for the listener) to suppose that Burton is at home.

I should point out that this notation is similar in many respects to linguistic notation, and purposely so. First of all, Burton isn't at home is an embedded structure. The proposition Burton is at home is embedded in another proposition that denies its truth, forming something that could be paraphrased as That Burton is at home is false; this is approximately the way the linguist would represent the negative (cf., e.g., McCawley, 1968). But what I have added to the notation is the notion of presupposition--what the listener is presupposing in uttering the denial. As we will see, it is this added notion of presupposition that is critical for differentiating the various types of negation and how they are comprehended.

Four classes of negation

We can, of course, ask the same question of all the other types of negation: What is it that people know of the meaning of a sentence once they have "comprehended" it? Since this step will be crucial to the "true" model of negation, it is necessary for us to agree on a representation for the various types of negation and to understand how these representations differ from each other. The most convenient way I have found for distinguishing the various types of negation from each other is to use two contrasts. First, negatives differ as to whether they apply to full propositions or to quantifiers. For example, absent is the negative of present and somehow implies a complete denial of the presence of something; on the other hand, few is the negative counterpart of many, yet few does not imply a complete denial of the presence of the number indicated by many. And second, negatives differ as to whether they are "explicit" or "implicit." Not present is an explicit negative, whereas its near synonym absent is an implicit negative. I will first discuss these two criteria and then detail the four categories that are formed by these two two-way distinctions.

Explicit vs implicit negation. Klima (1964) has pointed out that what I will call "explicit" negatives have several important syntactic consequences. Two such consequences are any-acceptance and either-conjunction. Sentences that contain an explicit negative are always able to accept the quantifier *any* in place of *some*. Consider the difference between *not present*, which is explicit, and *absent*, which is implicit, as shown in (2):

(2) a. John hasn't been present at any parties this month.
b. *John has been absent at any parties this month.

Though *John isn't present* and *John is absent* refer to the same situation, they differ in whether they accept *any*. Likewise, a sentence that contains an explicit negative always takes *either*, not *too*, as a final particle when it is the second of two conjoined sentences. Thus, in (3) and (4):

(3) a. *Max wasn't there and Minnie wasn't present too.
b. Max wasn't there and Minnie wasn't present either.
(4) a. Max wasn't there and Minnie was absent too.
b. *Max wasn't there and Minnie was absent either.

not present in the second of two conjoined clauses is tagged with *either*, whereas *absent* in the identical circumstances must be tagged instead with *too* indicating a positive second clause. So these two criteria--*any*-acceptance and *either*-conjunction--will serve to differentiate explicit from implicit negation on formal grounds. We will also have need to differentiate them on more specific semantic grounds.

Full vs quantifier negation. This distinction has a less formal basis than the explicit-implicit distinction. It might best be characterized as a difference between negation on dimensions with just two values, and negation on dimensions that have more than two values or are continuous. *Present-absent*, for example, is a positive-negative pair which defines the two values of the dimension (presence) that underlies it. In contrast, *many-few* is a pair which defines only two points on a quantified dimension of numerosity.

The explicit not is almost always a full negative, as far as I can tell, although not all explicit negatives are full, as we shall see. Perhaps the distinction that I am trying to make here will become clearer in the examples that follow.

1. Explicit full negation. This is the category of negation that is most commonly called negation. It includes sentence negation (Mort isn't at home), word negation (The car is polka-dotted, not striped), certain negative words (John forgot to bring any candy along), and so on. Little more needs to be said about this category.

2. Explicit quantifier negation. There are many examples of words that accept any and either, yet are not one of the commonly accepted negative particles. Few, for example, accepts any (Few people ate any of the avocados) and either (No men sang and few women sang either); hardly any and scarcely any have the same properties. The problem here, though, is how to represent adequately what people understand these explicit quantifier negatives to mean. Consider the sentence Few women sang. This sentence would be used in order to deny the supposition that the women that sang were many. That is, it might be represented by the notation in (5):

(5) a. Few women sang.

b. (false (suppose (women (women sang) were many)))

c. It is false to suppose that the women who sang were many.

Unfortunately, this notation incorrectly implies that few is identical in meaning to not many, and this is clearly not so. But in spite of the shortcoming of this notation, I will continue to use it, since it does indicate the other relations within the sentence in a concise way.

3. Implicit full negation. It is when we get into the implicit negations that we begin to see the importance of presuppositions in our characterization of negation. Consider absent, an implicit full negation vis-a-vis not present, an explicit full negation. Pat isn't present and Pat is absent clearly have the same range of denotations--they indicate exactly the same referential situation. Where, then, do they differ? The answer is, in their presuppositions. Note that Pat isn't present assumes that the listener has supposed that Pat is present, and it denies that supposition. In contrast, Pat is absent assumes no such supposition on the part of the listener. In this case, the sentence is affirming the proposition that Pat is not present, something the listener already supposes to be true or knows nothing about. Thus, the two sentences might be represented as (6b) and 7b), respectively:

(6) a. Pat isn't present.

b. (false (suppose (Pat is present)))

c. It is false to suppose that Pat is present.

(7) a. Pat is absent.

b. (true (suppose (false (Pat is present))))

c. It is correct to suppose that it is false that Pat is present.

The difference between the notations in (6) and (7) is obvious. Whereas it is the supposition that is being denied in (6), it is the content of the supposition that is negative in (7). Other examples of implicit full negation are the second member in the following pairs: to-from, in-out, on-off, etc.

4. Implicit quantifier negation. The last category of negation contains words like small, shallow, short, and most other adjectives that are "marked" (cf. Clark, 1969) or have negative force on their dimensions. Large and small,

for example, are not contradictories, like present and absent are, but rather are contraries; that is, not large does not necessarily imply small, although not present always implies absent. There is some question whether this category of words should even be called negative, although I will show that in many respects these negatives have psychological properties in common with the other types of negatives.

An interesting minimal pair of negatives can be found in few versus a few. While few is explicitly negative, a few is only implicitly so. Thus, few accepts any (Few people ate any rutabagas), while a few does not (*A few people ate any rutabagas); and few accepts either (No men sang and few women did either), while a few does not (*No men sang and a few women did either). But what is most important for our purposes is that few and a few can be shown to differ in their presuppositions. A few people left is an affirmation, while Few people left is a denial that many people left. This can be seen in (8) and (9):

(8) a. I thought everyone would leave, but few did.
    b. *I thought no one would leave, but few did.
(9) a. *I thought everyone would leave, but a few did.
    b. I thought no one would leave, but a few did.

These examples show that few can be used to deny a positive expectation, as expressed in the first clause of (8a), but it cannot be used to affirm an already negative supposition as in (8b); exactly the opposite is true for a few in (9). In short, a few is implicitly negative in the sense that it too is a negative in contrast with many, and it implies that the supposition itself is negative.

The problem here is how to construct a representation for few and a few that will show their differences. The tentative solution is to treat them just as we did not present and absent. The representations in (10) and (11) will suit this purpose:

(10) a. Few people left.

b. (false (suppose (people (people left) were many)))

c. It is false to suppose that there were many people who left.

(11) a. A few people left.

b. (true (suppose (false (people (people left) were many))))

c. It is correct to suppose that there weren't many people who left.

Scope of negation

In explicating the four types of negation--and there are probably other subtler distinctions to be made in negation as well--, I have noted two main things: (1) explicit negations deny what is supposed by the listener, while implicit negations affirm the listener's supposition, which is itself negative; (2) negations can be either full (as in present-absent) or quantifier (as in many-few). But there is another informal way of talking about these distinctions, and that is by talking about scope of negation. Note that in explicit negation, the scope of the negation is always larger than in implicit negation. The false in sentence (10) includes within its scope--those embedded strings that follow it--more than the false on (11). One of the hypotheses that I will discuss later on is that the difficulty of a negative is directly related to its scope: the greater the scope of the negative, the harder it will be.

The model of negation

William Chase, of Carnegie-Mellon University, and I have recently proposed a model for the comprehension of negative sentences in the context of a verification task (cf. Clark & Chase, in preparation). The most direct way to present this model is to describe it in relation to the specific task we used, and later I will show that this model is more general in that it can account for the comprehension of negation in a variety of tasks.

The Ss in Clark & Chase were presented displays that contained a sentence on the left and a picture on the right, and they were required to decide as quickly as possible whether the sentence was true or false of the picture. The Ss, then, were timed from the moment the display was visible to them up to the moment that they pressed either the "true" or "false" button with one of their thumbs. The problem we set ourselves was to account for the verification latencies--the total time between the beginning of the display and the response.

For this purpose, we assumed that the total process could be broken down into four stages. At Stage 1, the S encodes the sentence in a mental representation like those I have given in the previous section. At Stage 2, the S encodes the picture in the same sort of representational format. At Stage 3, the S compares the representation he has constructed for the sentence with that he has constructed for the picture to see whether they match or not. And at Stage 4, he takes the output of this comparison stage and converts it into some sort of a response. To make a model like this work, of course, one must make a number of assumptions about the process--e.g., what the sentence representations look like, how the pictures are encoded, what the comparison process is, and so on. Since the justification of this model and its assumptions is a monumental job in itself and has been fully described before (Clark & Chase, in preparation), I will here only give the conclusions of that study.

The sentences we used included sentences like The star is above the plus, The star isn't above the plus, etc., and the pictures on the right of the displays were a star (a typed asterisk) either above or below a plus. To simplify things, let me use simply the four schematic sentences A is above B, B is above A, B isn't above A, and A isn't above B--the original experiments also contained the comparable sentences with below--and the schematic picture of an A above a B. The first two sentences are positive, while the second two are negative, and the first and third are true of the picture, while the second and fourth are false. At Stage 1, the model proposes that these four sentences are represented as in (12b) through (15b):

(12) a. True Positive: A is above B.
b. (A above B)
(13) a. False Positive: B is above A.
b. (B above A)
(14) a. True Negative: B isn't above A.
b. (false (B above A))
(15) a. FAlse Negative: A isn't above B.
b. (false (A above B))

(For the present, I will omit the suppositional part of the negative for purposes of simplification.)

At Stage 2, the picture is assumed to be invariably coded as (A above B). Even though this is a gross simplification of actual coding process (cf. Clark & Chase, in preparation), this assumption is accurate for purposes of illustration.

The comparison stage

The Stage 3 comparison process, however, is the crucial part of the model of negation that we have proposed. The comparison process is built on the assumption that the representation of the sentence is compared to that of the picture in a series of steps. The principle underlying this process is that the S seeks to make the two representations exactly identical. This has been stated previously (Clark, 1969) as the principle of congruence: Ss cannot retrieve linguistic information from memory unless the representation of the information sought is completely congruent with the representation of the question asked of this information.

The whole purpose of the comparison stage in the verification task is to decide whether the sentence is true or false of the picture. For this purpose, Stage 3 is said to keep track of a truth index. This index starts out with the value true in it, and when some comparison operation of Stage 3 discovers a mismatch between the sentence and picture representations, another operation might change the value of the index from true to false, or perhaps later from false back to true again, depending on the circumstances. The final value of the truth index is the output of Stage 3 that will serve as the basis for the response that Stage 4 will have to make. If the index's final value is true, the S will respond at Stage 4 by pressing the "true" button; if the index reads false, Stage 4 will effect the pressing of the "false" button. Given these preliminaries, we can now turn to the actual proposal for the Stage 3 comparison operations.

For the four sentences listed in (12) through (15) and the picture encoded as (A above B), Stage 3 consists of a series of four ordered comparison operations. These operations can be visualized as a small computer program, as a flow diagram, as a list of branching rules, or whatever; in any case, the operations constitute an algorithmic method for coming to the correct truth value of the sentences. I will list these operations as a set of branching rules in Table 1. Consider, for example, the True Negative sentence B isn't above A, as represented in (16a), to be compared against the picture of an A above a B, as represented in (16b):

(16) a. (false (B above A))

b. (A above B)

Operation 1 would compare (B above A) from the sentence with (A above B) from the picture--(16a) and (16b), respectively--, and it would find a mismatch. So the process would go on to Operation 1a, which would change the current value of the truth index (presupposed at the beginning to be true) from true to false. Then Operation 2 would compare the (false()) of the sentence against the (())--i.e., no embedding string at all--of the picture, find a mismatch, and then give the control to Operation 2a. Operation 2a would change the current value of the truth index--now false because of Operation 1a--from false back to true. The final value of the truth value is therefore true, so Stage 4 would execute the push of the "true" button. When the same four operations are applied to the other sentences in (12) through (15), they produce the correct truth value in every case.

The importance of this model is that with only a few simple assumptions it is able to predict the verification latencies of the four types of sentences with extremely high accuracy. The assumptions are the following. First, Stages 1 through 4, and the four operations of Stage 3, are carried out serially, not in parallel. Second, the times consumed by each stage and each operation are constant from one condition to the next. So finally, the times consumed by the separate stages and operations are additive. At Stage 1, then, it is assumed that positive and negative sentences take a certain amount of time to set up, but that negatives take an increment of time b longer to set up than positives. At Stage 2, the picture always takes the same amount of time to set up. At Stage 3, Operations 1, 1a, 2, and 2a are assumed to take x, c, y, and d increments of time to perform, respectively. At Stage 4, it is assumed that "true" and "false" responses take equal increments of time to perform. The increments x and y at Stage 3, however, do not differentiate among the four sentences in (12) through (15), since each of the sentences requires both Operations 1 and 2. In summary, it is evident that the four sentences differ only in the increments of time b, c, and d, as shown in Table 1. False positives, for example, should be an amount c slower than true positives, whereas false negatives should be an amount c faster than true negatives. The parameters b and d are perfectly correlated, so they are listed as if they were a single parameter (b + d). The time consumed by all the other processes these sentences have in common is given by t_0, the wastebasket parameter. Thus, the four sentences differ on two parameters: c, which I will call Falsification Time; and (b + d), which I will call Negation Time.

Insert Table 2 about here.

Experimental evidence

Several experiments are reported in Clark & Chase (in preparation) and Chase & Clark (in preparation), all of which support the model just presented--the "true" model of negation. To give an example, in one experiment in which the Ss always looked at the sentence before the picture, Falsification Time c was estimated to be 187 msec and Negation Time (b + d), 685 msec. These two parameters (along with another one that is independent of these two for the sentences containing below) accounted for 99.6% of the variance in the means of the four conditions with either above or below as the preposition and a median of 95% of the variance for each S considered separately. Note that the "true" model predicts that True Positives will be the same amount of time c faster than False Positives as True Negatives are slower than False Negatives. This prediction was confirmed to the millisecond in this experiment. The other experiments, which were carried out with other important experimental differences, all confirmed this model too. Across four different experiments, for example, Falsification Time c was estimated to be 187, 148, 145, and 145 msec, respectively, and Negation Time (b + d), to be 685, 556, 660, and 660 msec, respectively; this shows the great consistency of these estimates across a wide variety of experimental conditions.

But perhaps just as important as the verification latencies are the reports Ss make as to what they think they do in the task. They in fact report something quite comparable to the series of stages and operations that are given in the "true" model. Concerning the Stage 3 operations, they point out that on positive sentences they check to see if the sentence "agrees" with the picture or not; if it does not they make their answer false--otherwise, they leave it true. But for the False Negative sentence A isn't above B, for example, they note that the picture shows a picture of an A above a B, whereas the sentence claims the A is not above the B; so the sentence is contradictory, and they change their answer to false. On the true negative B isn't above A, the Ss note that the B is above the A and this does not agree with the sentence; but since the sentence is negative, it must be true anyway. In other words, the Ss often report "changing their answer" once for False Positives and False Negatives, and changing it twice for True Negatives.

Wason (1961) has also reported similar intuitions on the part of his Ss. Since these reports agree precisely with what the model says that Ss do, the model is further justified by these subjective reports.

The "conversion" model of negation

Although almost all of our Ss have reported using this two-step comparison process, Wason (1961) and others have reported that some Ss say that they "convert" each negative sentence into a positive one before they ever try to verify it. Wason (1961), for example, reported that Ss often said they converted Six is not even into Six is odd before trying to decide whether the sentence was true or false. But Young & Chase (in preparation), in a very important series of experiments, showed that Ss can be instructed to carry out conversions such as this one; nevertheless, the results are still well accounted for by the "true" model of negation. But let me amplify.

In one of Young & Chase's experiments, Ss were asked to convert negative sentences into positives, whenever they met up with one, by the following rule: Change X isn't above Y into Y is above X. (Again for simplification, I will ignore all the sentences containing below.) This instruction implies that instead of representing the True Negative B isn't above A directly as (false (B above A)), they "converted" it first and represented it at Stage 1 as (A above B). Likewise, they represented the False Negative, not as (false (A above B)), but as (B above A). In their final representations, of course, the True Positive and True Negative are identical, and so are the False Positive and False Negative. At Stage 3, then, the "true" model would predict that both positives and negatives should require just one comparison operation, Operation 1. Operation 2--and its partner Operation 2a--will never be necessary. To predict verification latencies, Young & Chase assumed (1) each conversion took a constant amount of time, and (2) otherwise the "true" model of negation held. By these assumptions, negatives should still be slower than positives, but now a true sentence should be faster than a false one regardless of whether the sentence is positive or negative. Young & Chase confirmed these predictions very accurately. And they confirmed the analogous predictions for the "conversion" rule to change isn't above into below and isn't below into above, and for two other more complicated "conversion" rules. So their experiments constitute an excellent demonstration that the encoding and comparison stages of the task are quite separate, and that the comparison stage of the "true" model is consistent with their data.

It is important here to contrast the "true" and "conversion" models of negation. Both models predict that negatives will take longer to verify than positives. In the "true" model, this is because of the representation time $\underline{b}$ plus the Operation 2a time $\underline{d}$ that negatives require over positives. But in the "conversion" model, this is because the conversion itself takes time to carry out, an increment we might call the Conversion Time $\underline{k}$. (Interestingly enough, Young & Chase found that $\underline{k}$ was often less than negation time ($\underline{b} + \underline{d}$) for the identical problems under conversion vs. non-conversion instructions.) But the significant difference between the "true" and "conversion" models is that true-false should interact with positive-negative in the "true" model, but not in the "conversion" model. Specifically, under both models True Positives should be faster by an amount $\underline{c}$ than False Positives. Yet True Negatives should be slower by $\underline{c}$ than False Negatives when Ss do not convert, but faster by $\underline{c}$ when Ss do convert (regardless of the conversion). When Ss are using the "true" method, then, true sentences should be no faster overall than false ones, although when they "convert," true sentences should be faster by c overall than false ones.

By now it should be clear why I have chosen to call these two methods the "true" and "conversion" methods of negation. The "true" method works for all negatives in every task, and it treats negatives--i.e., it represents negatives--with all the negative information in its proper form. The "conversion" method is, in a sense, a way of cheating in comprehending the negatives. It changes negatives into positives and does not retain all the negative information in its proper form. To see how this is cheating, consider A isn't above B to be compared to a picture of an A and a B side by side. If the S "converts" A isn't above B, say, into A is below B, he will look at the A and B side by side and claim that the sentence is false of the picture. But this claim is incorrect just because the positive sentence after "conversion" is not synonymous with the negative one before "conversion." Of course, the S would have made the correct judgment if he had used the "true" method, since a comparison of (false (A above B)) with (A beside B) would have correctly led him to say "true" under the Stage 3 comparison rules in Table 1.

Previous Studies of Explicit Negation

We can now turn to the extensive literature on the comprehension of negation. The purpose of this section is simply to demonstrate that the "true" model of negation, with its illegitimate offspring the "conversion" model, is able to account for the main variations in this previous literature. The first studies to be examined are all verification tasks in which the S had to judge whether a sentence was true or false. In a second set of studies by Jones (1966a, b, 1967), the negative is found in an instruction the Ss were to follow.

The verification studies

Beginning with Wason's study in 1961, many psychologists have asked Ss to verify sentences, but they have relied on three different types of verifying information. Sentences were sometimes verified against previous knowledge, as when Ss verified Six isn't odd; sometimes against pictures, as in Clark & Chase (in preparation); and sometimes against other sentences, as in X doesn't precede y, therefore y follows x (Greene, 1970a). To be able to generalize across these differing tasks (and their highly similar results), it is therefore necessary that we represent the verifying evidence --whether it be previous knowledge, a picture, or another sentence--in the same format. This is essentially the assumption that was made in Clark & Chase in order to derive predictions for verifying pictures, and it was well demonstrated in other ways in that study. So I will make the assumption that all three types of verifying evidence are represented in a semantic representation of the sort I have given above for sentences and pictures.

Consider Wason's experiments in which he asked Ss to verify sentences like Six is not odd. At Stage 1, the S would represent Six is not odd as (false (6 is odd)); at Stage 2, he would represent his previous knowledge about six as (6 is even); at Stage 3, he would compare the two representations in the four-operation process of Table 1; and at Stage 4, he would produce the response. But Wason, in questioning his Ss, found that about half of them reported making conversions, changing not even to odd and not odd to even, whereas the other half appeared to use what I have called the "true" method.

In other words, Wason's results should show a mixture of these two types of Ss. First of all, True Positives should be faster than False Positives for both types of Ss, and this is in agreement with Wason's results. But second, True Negatives should be faster than False Negatives for the "converters," but slower for the non-cheaters. Overall, this should make True and False Negatives about equal in verification latency, and this is exactly what Wason found. So Wason's results, although not in agreement with the "true" model taken alone, do agree with the "true" model when account is taken of those Ss who consistently made "conversions."

A perusal of the other studies on explicit negation and its verification shows that all the other studies (Eifermann, 1961; Wason & Jones, 1963; Gough, 1965, 1966; Slobin, 1966; Wales & Grieve, 1969; Greene, 1970a, b; Trabasso, Rollins, & Shaughnessy, in press) also succumb to this analysis (cf. Clark, in press). In some instances, the "conversion" method was impossible to use, and in those cases, the "true" model of negation fits the data very well; in the other cases, where there was reason to think that some Ss were "converting," the results look like a mixture of the two methods, just as Wason's results did. Trabasso et al. (in press) were the most explicit on this point. They used the identical verification task in two conditions: in one, Ss could convert (and reported doing so) and the results followed the "conversion" model almost precisely; in the other, "conversion" was literally impossible for methodological reasons, and the results followed the "true" model, just as they should. The previous literature on verification tasks, therefore, constitutes good support for the "true" and "conversion" models that have been proposed here.

Following instructions

The verification task is just one place where negative sentences must be comprehended. Another obvious place is in tasks in which the S must carry out an action in conformity to an instruction, and the instruction can be given in either positive or negative terms. Indeed, that the "true" model of negation is just as applicable to this kind of task is demonstrated in a very interesting series of experiments by Jones (1966a, b; 1967).

In one experiment, Jones gave her Ss either the positive instruction in (17a) or the negative one in (18a); the respective representations for them are approximated by (17b) and 18b):

(17) a. Mark the numbers 3, 4, 7, 8.

b. (you mark X (if (X is (3 or 4 or 7 or 8))))

(18) a. Mark all numbers except 1, 2, 5, 6.

b. (you mark X (if (false (X is (1 or 2 or 5 or 6)))))

The Ss were given a page full of numbers--random numbers selected out of the digits 1 through 8--and one of the two instructions, and were told to cross out all the appropriate numbers on the page as quickly as possible. They were timed. The basic task can be viewed as follows. At Stage 1, the Ss represent the instruction appropriately as in (17b) or (18b). At Stage 2, they represent the number they are currently looking at (say 7) in an implicit question like "Should I cross out 7?", which would be represented as (you mark X (if (X is 7))). At Stage 3, the Ss would compare the if-clauses of the two representations by the four-operation process in Table 1. And by Stage 4, there is an answer to the implicit question--either (true (you mark X)) or (false (you mark X))--which is then translated into a response of either marking out the 7 or not.

Conceived of in this way, Jones' task should fit predictions from the "true" model of negation. The predictions are relatively simple. Instruction (17) should take less time than (18), since the if-clause is positive in (17) but negative in (18). Jones' results are in agreement with this prediction. Unfortunately, we cannot test for latency differences in True and False Positives and True and False Negatives, for her data were not so collected. But we can test the same predictions on the errors Jones' Ss made. We must first assume that Ss should make more False Positive than True Positive errors, since a False Positive decision requires one more mental operation (Operation 1a in Table 1) than a True Positive decision and that operation might be missed. For precisely the same reason, Ss should make more True Negative than False Negative errors. Jones' results confirm both of these predictions. For the positive instruction (17), she found that Ss made 2% True Positive errors (crossing out digits they should not have), On the negative instruction (18), the pattern was just the reverse: Ss made 8% True Negative errors (crossing out digits they should not have), but only 4% False Negative errors (failing to cross out digits they should have).

In summary, Jones' data on positive and negative instructions--and I have only discussed one of her experiments here (cf. Clark, in press and below)--fully support the hypothesis that her Ss understood these instructions consistent with the "true" model of negation.

Comparative studies of negation

After this brief review of the literature on explicit negatives, I now turn to several recent comparative studies of negation. These studies are comparative in the sense that they were specifically designed to compare two or more contrasting types of negation.

In the past, investigators have stuck very close to the explicit full negative, and as a result, the previous studies appear to emphasize the explicit negative as something special. Indeed, some of these studies have given special attention to the syntactic transformation that inserts the negative particle into a positive sentence as if it was the syntactic properties of the negative that gave it its properties in comprehension. So one important conclusion to be drawn from the following studies is that explicit and implicit negatives constitute a unified notion of negation as far as comprehension is concerned. Negation is fundamentally a semantic notion. Although negatives are sometimes expressed syntactically and sometimes lexically, it is the underlying semantic representation that is important for comprehension.

A second important conclusion to be drawn from the studies to be described is that scope of negation and presupposition are intimately involved in the comprehension of negatives. As we will see, certain critical differences in comprehension seem to depend solely on the fact that two negatives differ in their presuppositions or in the scope of their negative marker.

Present and absent

The first study I know of to compare explicit and implicit negation is found in Clark & Young (in preparation), in which sentences containing present, absent, and their negatives were verified against the presence or absence of objects in pictures. Present and absent were chosen because The plus isn't present and The plus is absent both have the same truth value and require the same response for particular pictures; in other words, the two sentences are experimentally comparable. The expectation was that The *plus isn't present* would be verified by the "true" method of negation, so that if *The plus is absent* is also a negative, it should behave in exactly the same way.

These expectations were verified very nicely, with one important exception. True sentences containing *not present* and *absent* were verified more slowly, respectively, than false sentences containing *not present* and *absent*; this is what the "true" method of negation would predict. But even though the *absent* sentences were completely parallel to the *not present* sentences, the former took much less time to verify. Specifically, the negation time (*b* + *d*) was about 371 msec for *absent*, but about 640 msec for *not present*. This suggests, of course, that *The plus is absent* and *The plus isn't present* cannot be represented in the same way, or else they should show approximately equal negation times.

But as was pointed out previously, *isn't present* and *absent* make different presuppositions about the underlying proposition they are negating. *The plus isn't present* should be represented as (*false (suppose (plus present*))), whereas *The plus is absent* should be represented approximately as (*true (suppose (false (plus present*)))). The difference in negation time between *not present* and *absent* should therefore probably be related to the difference in their representations. The question becomes, then, does the difference lie in (1) the comparative times it takes to represent *absent* and *not present* at Stage 1, or (2) the comparative times consumed by the Stage 3 comparison operations? Although very little evidence is now available to decide on this issue, one result of Jones'--to be discussed shortly--seems to suggest that (2) might be true. This would suggest that the *d* increment, the time taken by Operation 2a in the comparison process of (16), would simply be longer when the false dominates the supposition, as in not present, than when it is dominated by suppose, as in absent. It should be pointed out here, however, that the not present vs. absent difference of 271 msec could not be accounted for by the extra reading time of not present; on independent evidence, the extra n't could take no more than 50-100 msec to read.

Perhaps, then, it is wisest not to speculate on the mechanisms that underlie the not present vs. absent difference until more specific evidence can be found for the several alternatives. Rather, we might simply state it as a hypothesis about scope of negation, one that is to be accounted for later. The hypothesis is this: the greater the scope of negation, the longer it takes to comprehend the negative. Fortunately, the correctness of this hypothesis is not directly relevant to the correctness of the "true" model of negation. It might be that the specific form of the rules in Table 1 will have to be modified in its details, but the results of Clark & Young make it necessary for the basic two-comparison verification process to hold in some form or other in all the negatives examined so far.

"Except" revisited

The scope of negation hypothesis receives further support in a study by Jones (1967), if certain assumptions are made about the representations of the sentences she used in her study. In this study, like the other one reported on above, Jones gave Ss one of two instructions about crossing out digits, gave them a page full of digits, and timed them as they crossed out the appropriate digits on each page. But in this case, the instructions she compared (among others) were (19a) and (20a), which I presume are represented as in (19b) and (20b), respectively:

(19) a. Mark all the numbers except 2, 5, 8.

b. (you mark X (if (false (X is (2 or 5 or 8)))))

(20) a. Mark all the numbers, but not 2, 5, 8.

b. (false (you mark X (if (X is (2 or 5 or 8)))))

That is, the scope of the except pertains only to the testing of the numbers themselves in (19), but the scope of false in (20) covers the whole instruction. Both instructions, of course, would have the same result by the four-operation process of Table 1: it is just that the false enters into the process in different places for the two instructions. Thus, if the scope of negation hypothesis is correct, then instruction (19) should be easier than (20). Jones' results support this prediction. She also found that her Ss often restated the instructions to themselves before carriing out the task, and they most often ended up using instruction (19) instead of (20), as if (19) was the easier instruction to use.

If this interpretation of Jones' results is correct, then her results imply that the greater difficulty with greater scope of negation must be accounted for, at least in part, during the Stage 3 comparison process. Presumably, the Ss did not have to re-encode the instruction each time they crossed out a new digit. Since the only process that recurred for each digit was the comparison process, the source of increased difficulty of instruction (20) over (19) must lie in the comparison stage.

The Just and Carpenter study

Very recently, two students, Marcel Just and Patricia Carpenter, have completed a comparative study of several types of negation in English. They asked Ss to verify sentences like Most of the dots are black and A small proportion of the dots are black against pictures that contained either 14 black dots and 2 red dots or the reverse. They explicitly set out to study three categories of negation. Category 1 contained three examples of explicit negation: (A) The dots are red vs. The dots aren't red; (B) There are red dots vs. There are no red dots; and (C) All of the dots are red vs. None of the dots are red. (The pictures paired with the sentences of Category 1, unlike those for 2 and 3, contained either all red dots or all black dots.) According to the present classification, Example A is an explicit full negation, while B and C are both explicit quantifier negations, although the latter classification is debatable. Category 2 contained examples of explicit quantifier negation: (D) Most of the dots are red vs. Hardly any of the dots are red; (E) Lots of the dots are red vs. Scarcely any of the dots are red; and (F) Many of the dots are red vs. Few of the dots are red. Finally Category 3 contained three examples of implicit quantifier negation: (G) The majority of the dots are red vs. The minority of the dots are red; (H) A large proportion of the dots are red vs. A small proportion of the dots are red; and (I) 14 out of 16 dots are red vs. 2 out of 16 dots are red.

Their results support the "true" model of negation in several very important ways. First, they found that negatives were more difficult than affirmatives for all three categories. Second, they found that in Categories 1 and 2, which contained all explicit negatives, True Positives were faster than False Positives, and that True Negatives were slower than False Negatives. This agrees perfectly with the "true" model of negation. But third, they found, quite unexpectedly, that in Category 3 True Positives and True Negatives were both verified faster, respectively, than False Positives and False Negatives. One might conclude that unless the Ss were carrying out some sort of "conversion" while encoding the sentences, the "true" model of negation should not expect this result. But Just & Carpenter explained this apparent inconsistency in quite a different way.

Their explanation of the Category 3 inconsistency lay in how the Ss coded the pictures. (The terminology in the explanation to follow, however, is mine.) Consider many-few (a Category 2 pair of quantifiers) vs. many-a few (substituting for a Category 3 pair of quantifiers). Given the true sentence Many of the dots are black--i.e., (true (suppose (dots (dots black) are many))) --, the S could code the picture in terms of the majority of dots, as (suppose (dots (dots black) are many)), or in terms of the minority, as (suppose (false (dots (dots red) are many))). Note that the coding of the majority of dots will be congruent with the sentence, but the coding of the minority will not. Just and Carpenter simply assumed that Ss coded the majority of dots for both many and few (of Category 2), and this would produce the correct predictions from the "true" model of negation. In contrast, they assumed that Ss coded the majority and minority, respectively, for many and a few (of Category 3); this likewise allows the correct predictions from the "true" model. To put it another way, Ss were always assumed to code the picture in terms of the presupposition of the sentence: many and few both contain suppositions about the majority of the dots; but many and a few contain suppositions about the majority and minority of the dots, respectively. And this coding is natural. Many and few respectively affirm and deny something about the majority of the dots; on the other hand, many and a few are both affirmations, one about the majority of the dots and the other about the minority.

Just and Carpenter were aware that their explanation needed further justification. So they re-ran Categories 2 and 3 of their first experiment, but this time they presented the pictures a half-second before the sentence appeared. In one condition, the Ss were asked to code the majority of the dots from the picture, and in another, the minority. Their results were in full agreement with their previous explanation. When Ss coded the majority of the dots, Categories 2 and 3 were in complete agreement, in contrast with the results of their first experiment. And when Ss coded the minority, the two categories were again in complete agreement, but in this case True Positives and False Negatives were slower, not faster, than False Positives and True Negatives, respectively. This result is also consistent with the "true" model of negation, once the newly instructed codings of the pictures have been taken into account. Therefore, when Ss are forced to code the pictures in one particular way, the Ss then verify the exemplars of Categories 2 and 3 in exactly the same way. So the Just & Carpenter results fit nicely with the present unified explanation for negation.

The Just & Carpenter results also add evidence for the scope of negation hypothesis. Their exemplars of negation can be approximately ordered in terms of decreasing scope. Exemplar A is sentence negation and should have roughly the greatest scope; Exemplars B, C, D, E, and F are all explicit negatives whose scope include the presupposition, so they are next; Exemplars G, H, and I have the least scope, since their negatives are included within the supposition. Significantly, the negation time of these three groups of exemplars decreased approximately with scope. This evidence, however, is not as direct as in the present-absent and except examples discussed above, since so many other things seem to be varying at the same time in Just & Carpenter's exemplars.

Other studies of implicit negation

Although I know of no other strictly comparative studies of negation, there are a number of studies that have investigated the comprehension of other implicit negatives. These studies fall into two categories--those on adjectives, and those on prepositions.

Adjectives. We have already discussed present-absent as a positive-negative pair. Although there are no other studies on adjectives that might be considered full negatives, there are a series of studies on implicit quantifier negatives that I have reviewed recently (Clark, 1969) in relation to deductive reasoning. The main generalization I drew there was that positive adjectives take less time to comprehend than negative ones, at least when they are found in comparative form. Thus, deductive reasoning problems that contained the comparatives of good, much, fast, far, tall, happy, warm, deep, old, and high were solved consistently faster than problems with their respective opposites. This evidence, along with the results of Just & Carpenter, lead us to the tentative conclusion that all the lexically marked adjectives like bad, little, slow, etc., are implicit (quantifier) negatives.

Prepositions. Elsewhere (Clark, in press), I have argued that many pairs of prepositions can also be considered positive and negative. Gruber (1965), for one, has shown that in-out, on-off, and to-from are genuine positive-negative pairs. These negatives, out, off, and from, are full implicit negatives, since, for example, John is out of the house implies John isn't in the house, and vice versa. In and out, on and off, and to and from specify the only two semantic values that can be taken on their respectively semantic dimensions.

Carol Offir and I have recently carried out a couple of studies in which Ss were asked to verify sentences like It is going to town B against a picture of a symbolic car going to or from town A or B, or to read sentences like John thought, "Mary is about to go into the kitchen" and verify other consequent sentences like Mary could be in the kitchen. Of interest here are the results of these studies with respect to to and from, and into and out of. First of all, sentences with to were verified more quickly than those with from, and the same held for into and out of, respectively. In the second study, in which Ss verified the conclusion drawn about a previous sentence, it was found that out of behaved according to the "true" model of negation. (The full description of this evidence is too complex to give here.) So these two studies add to the growing evidence that the "true" method of negation seems to apply to implicit as well as explicit negatives.

In Clark (in press), it is also argued that pairs like above-below, ahead of-behind, in front of-in back of, and other such pairs are also implicitly positive and negative. The negative in these pairs, however, is quantifier rather than full negation, since The red balloon isn't above the blue balloon does not necessarily imply The red balloon is below the blue balloon, a fact which contrasts with the full negative case in which John isn't in the house always implies John is out of the house. In Clark (in press), I have reviewed a number of studies in the literature as well as some of my own that show that above, on top of, in front of, ahead of, and before are comprehended more quickly than their negative opposites below, under, in back of, behind, and after.

In summary, the evidence for the comprehension of certain implicitly negative adjectives and prepositions are in general agreement with the "true" model of negation.

Recapitulation of the comparative studies

We can now bring together some of the diverse evidence that supports the scope of negation hypothesis. In Tables 3 through 6 I have collected together the studies I know of that have measured the latencies in the verification of positive and negative sentences from the moment the sentence was presented to the moment the S responded. In each of the four tables, I have listed the investigators on the left, the negative studied in the center (the positive counterpart of which is usually obvious), and the Negation Time for each negative on the right. The four tables list the negation times, respectively, for explicit full negatives, explicit quantifier negatives, implicit full negatives and implicit quantifier negatives.

There are several observations that we can make on these tables. The most obvious is that Negation Time is positive regardless of the negative, although Negation Time varies from about a tenth of a second to almost three-quarters of a second, depending on the negative.

The second observation is that the median Negation Times order themselves from longest to shortest as follows: explicit full negation (600 msec), explicit quantifier negation (258 msec), implicit full negation (185 msec), and implicit quantifier negation (145 msec). That is, Negation Time decreases approximately as we go from explicit to implicit negation, and from full to quantifier negation. In more detail, it is clear that explicit full negation has the longest Negation Time; its exemplars, which vary from 440 msec to 736 msec, do not overlap with those of any other category. The exemplars of the other three categories have considerable overlap with one another, although there is at least a hint that the medians might be reflecting real differences among the latter three categories.

The importance of this last observation, if it is correct, is that the median Negation Times follow the scope of negation. Explicit negatives have greater scope of negation in general then implicit negatives, and full negatives have greater scope than quantifier negatives, and these differences are correlated with differences in Negation Time. These data, then, constitute support, albeit less than overwhelming support, for the scope of negation hypothesis.

A third observation that can be made is that the Negation Times are _relatively_ homogeneous within each table despite the wide variety of exemplars there. Take, for example, the explicit full negatives in Table 3 and their Negation Times, which vary from 440 to 736 msec. The exemplars in this table include locatives (_Star isn't above plus_), attributives (_The dots aren't red_), and predicate nominatives (_Seven is not an even number_). The exemplars also include sentences in which one word only--other than the negative--is significant for the verification (_The dots aren't red_), two words (_Seven is not an even number_), or three words (_Given 6 and 8, the next number is not 1_). In other words, it appears that Negation Time is around 600 msec regardless of other incidental properties of the sentences. Negation Time is as long as it is because sentences contain explicit full negations--in these instances sentence negations. If this observation is supported by future research it constitutes strong evidence for the independence of Negation Time from Falsification Time and from other attributes of sentences, and hence for the separability of the four Stage 3 comparison operations shown in Table 1.

Summary data like these, of course, should be interpreted with a great deal of caution. First, the experimental methods varied considerably from study to study; in some cases the experimental error is large, but in others it is small. Second, and probably more important, we have no idea how the other differences in the negative exemplars--like the difference between _from_ and _out of_--affect Negation Time as well. It seems only reasonable that there should be other differences among the negative exemplars. Third, Negation Time can be significantly affected by the range of objects (e.g., pictures) that the sentences are verified against. Consider, for example, the three instances of absent listed in Table 4. The first sentence was verified against a picture of a plus ("true") or a star ("false"), the second, against a picture of an empty square ("true") or a circle within a square ("false"), and the third, against a picture of a disk absent ("true") or present ("false"). The three corresponding Negation Times, then, vary because of subtle interactions between the negative sentences themselves and the type of evidence they are compared against. I will discuss the reasons for the differences in this particular example later, although in the other exemplars in Tables 3 through 6, I am quite uncertain as to which variations in experimental procedure are important and which are not.

One final note of caution. Young and Chase (in preparation) and Chase, Young, and Clark (in preparation) have run Ss for many more than one session. Although Negation Time was found to begin at around 600 msec for an explicit full negative, Negation Times diminished over the next 10 to 15 days to as little as 100 msec. In other words, Negation Time can be reduced with practice. The experiments listed in Tables 3 through 6 could also well differ in how practiced the Ss were--either on negatives in general or on the particular negatives of the experiment. So some of the variation in the tables--and one cannot easily tell which variation--might be partly due to practice effects.

The support for the scope of negation hypothesis, then, is still quite meager, though several pieces of evidence suggest that it might be correct. The most direct evidence is that Negation Time for absent is considerably less than for not present (Clark & Young, in preparation), and that the instruction "Mark all the numbers except 2, 5, 8" is easier than the instruction "Mark all the numbers but do not mark 2, 5, 8" (Jones, 1967). Suggestive, but not conclusive, is the combined evidence from the extant studies on negation, which shows that explicit full negation has longer Negation Time than explicit quantifier negation, implicit full negation, or implicit quantifier negation, and that the latter three categories might possibly be ordered in this way from longest to shortest Negation Time. As yet, there seem to be no counter-examples to the hypothesis. Full confirmation of this hypothesis waits on studies, like that of Clark & Young, in which scope of negation can be varied without concomitant changes in verifying evidence, practice, and other confounding effects.

Negation in cognition

To round out my sketch of the evidence for the "true" model of negation, I will now consider several examples from perception that appear to require the present model of negation. Although perception has traditionally been kept quite distinct from psycholinguistics, it is obvious that there must be a representational system common to perception and language, for otherwise it would be impossible for people to talk about pictures, to verify sentences against pictures, to imagine a scene as described by a sentence, and so on. Indeed, Chase and I have presented evidence (e.g., Clark and Chase, in preparation) that supports the position that, at some level of processing, pictures are represented in semantic representations that are just like those of sentences. For example, an A above a B is normally represented as (A above B), which is identical to the underlying representation of the sentence A is above B. It is necessary to make this assumption simply to explain a large variety of findings in the verification studies on sentences like A is above B, B isn't below A, etc. But the assumption is also necessary in order to bring together all the other studies on the verification of sentences. As pointed out before, if sentences, pictures, and prior knowledge were represented in entirely different forms, then it would be impossible to explain all the previous studies on negation with one unified theory, and the latter alternative seems highly preferable to three unrelated theories of comprehension. More evidence against the alternatives of treating pictures as pictures, etc., instead of as "interpreted" underlying semantic representations like (A above B) can be found in Clark & Chase (in preparation).

The main point of perceptual representations is that they are normally positive--that is, we code perceptual objects as positive entities. "That is a tree," we think, not "That isn't a bush" or "That isn't a rock" or "That isn't Julia" etc. The reason is clear. In most cases, there is only one possible identification of what a thing is, but many of what it is not. Yet there are instances where it is important or even necessary to think of something in negative terms. Significantly, the instances of negation in perception that I have found follow the same logic as the linguistic examples of negation discussed above; that is, negation in perception fits into the "true" model of negation, and it can be analyzed in terms of the presupposition the perceiver has of the situation he is observing. I will first discuss evidence for the normality of positive coding, followed by an example of a negative perceptual representation and its properties.

Positive perceptual coding

In Clark & Chase, it was shown that a picture of an A above a B can be "looked at" in at least two different ways. When the S is instructed to look at the top figure (A) in the picture, he will encode the picture as (A above B). When instructed to look at the bottom figure instead, he will encode the picture as (B below A). When simply asked to view the picture as a whole, the code is always (A above B). (The differences in these codings can be detected by the way the different viewing instructions interact with the verification of sentences like A is above B, B is below A, etc.) In other words, an A above a B is normally encoded in terms of the implicitly positive preposition above as (A above B), unless the S is instructed to view the picture in a specific way, and then the coding can become (B below A) instead. Here is the first piece of evidence for the proposition that pictures are encoded in positive representations except under unusual conditions.

A second finding of the same study is that the Ss could not have been coding the pictures as an explicit negative either. For example, there is no reason why the S could not have coded an A above a B as (false (B below A)), for this second coding would also be true of the picture. This negative coding, however, is also consistent with a large number of other sentences that would not be true of the picture, such sentences as A is beside B. In short, encoding a picture in an explicit negative representation would often lead to incorrect judgments of truth value, and there is evidence in the data of Clark & Chase anyway that Ss did not use such negative representations. In the present-absent study reported above, too, it was assumed that the pictures were encoded in a positive form. A picture of a plus was assumed to be coded as (plus is present), not as (false (star is present)). So here are two other examples of how pictures are normally encoded in their positive forms.

Negative perceptual coding

In order to overcome this very strong tendency to code perceptual information only in its positive form, I have recently carried out a verification task with pictures that are difficult to code in anything but a negative way. Consider a picture of a square with or without a smaller circle inside it.

142A

The square with a circle is meant to represent the presence of a hole, while the square without a circle is meant to represent the absence of a hole. The Ss were then asked to verify sentences in the usual manner: they were given sentences like The hole is absent together with a picture of a square with or without a circle inside it and were asked to judge whether the sentence was true or false of the picture. The sentences describing the presence or absence of the hole were mixed in with sentences describing the presence or absence of a star or plus; the pictures for the latter were either a plus or a star.

The rationale of the study was that the pictorial presence of the circle would be represented as (hole is present), while the pictorial absence could only be represented in a negative way, as (false (hole is present)). So instead of utilizing verifying and falsifying pictures that are both coded in positive form, this study utilizes pictures that are represented in positive and negative forms. The representations for the sentences describing the hole and for the corresponding pictures are shown at the top of Table 7, and they

---

Insert Table 7 about here.

---

are to be compared with the representations for the sentences describing the star or plus at the bottom of the table. When the four operation comparison process of Table 1 is applied to these representations, the process still produces the correct truth value in each instance, but the latency components for the hole-sentences turn out to be different from those for the star-sentences. In particular, note that c never appears in the hole-sentences, and b and d are not correlated in the hole-sentences as they are in the star-sentences. In short, the "true" model of negation predicts quite different patterns of latency for the two kinds of sentences.

The results of the experiment on the *hole*- and *star*-type sentences are in agreement with these predictions. For the sentences describing the star and plus, the results were as usual: True Positives and False Negatives were faster, respectively, than False Positives and True Negatives. But for the *hole*-sentences, True Positives and True Negatives were both faster than their fa counterparts: in this case there was no interaction between true-false and positive-negative. This is just what the "true" model would predict.

There is an important difference between the *star*-sentences and the *hole*-sentences. *The hole is absent* was always confirmed by the physical or perceptual absence of a figure in the picture (viz. by the absence of the circle within the square), whereas *The star is absent* was always confirmed by the physical presence of a figure (the plus). One might argue that the difference in latency patterns between the *hole*- and *star*-sentences is attributable to the perceptual properties, not to the semantic representation of the picture itself, as I have been arguing. But this possibility was eliminated in a carefully counterbalanced second condition in which *hole* was replaced by *lid*, so that *The lid is present* was true when the circle was *absent* from the square, and *The lid is absent* was true when the circle was *present*. As expected, the *lid*-sentences had exactly the same pattern as the *hole*-sentences. Thus, it is the semantic interpretation of the picture that is all important, not simply the physical attributes of the picture.

In summary, the *hole*- and *lid*-sentences constitute one example of how a picture must be interpreted negatively. Furthermore, the negative code depends crucially on the *Ss*' presuppositions about what the picture is supposed to be. If the picture is to be interpreted as a lid, one coding is used, but if it is to be interpreted as a hole, exactly the opposite coding is used.

When these representations are then used to verify a sentence, they behave exactly like ordinary linguistic positive and negative representations as predicted in the "true" model of negation. Thus, our perceptual system makes use of negation too, and this negation is fundamentally of the same sort that is found in language: it consists of a proposition (e.g., The hole is present) embedded within another predicate (it is false) that denies the truth of the embedded proposition.

Conclusions

In this paper, I have presented a partial theory of comprehension and have applied this theory specifically to negation. The general theory, as applied to the verification process, was simply that the process consisted of four stages: a sentence encoding stage, a picture encoding stage (if a picture is used as verifying evidence), a comparison stage, and a response stage. In applying this model to negation, I have argued that negative sentences and pictures are represented in a particular form and that the comparison stage carries out its duties in a particular way.

The representation of negation. The basic proposal is that negative sentences like Jeffrey isn't at home consist of two parts: (1) an embedded proposition, Jeffrey is at home; and (2) an embedding proposition, it is false, which denies the truth of the embedded proposition. A picture that it represented negatively will also take on the same form. Although there are many varieties of negatives in English, the proposal is that they are all represented in a variation of this basic form.

The comparison process. The "true" model of negation contains a comparison stage that consists of four ordered mental operations. Given a representation for a sentence and a verifying picture, the process first checks for synonymy of the embedded strings of the two representations, and then checks for the synonymy of the embedding strings. When there is a mismatch in either mental operation, there is another subsidiary operation that changes the truth index of the sentence--e.g., from true to false, or from false to true. It is the final value of the truth index that serves as the answer finally produced as a response. In the "conversion" model of negation, the S is able to avoid checking the embedding string during the comparison stage altogether simply by eliminating the embedding string when he encodes the sentence at the first stage. Thus, the S might immediately interpret Nine isn't even directly as Nine is odd, which he then represents without a falsifying embedding string. What I have tried to show is that all the existing evidence on negatives can be accounted for by either the "true" or the "conversion" models of negation or by some mixture of the two. In particular, these models account for (1) the comparative latencies in verifying True or False Positive or Negative sentences; (2) the introspective reports Ss give as to how they go about verifying sentences; and (3) the latencies and errors Ss make in following positive or negative instructions. To account for these results, I have had to assume that the comparison operations are serially ordered and that each operation consumes a given increment of time on each verification.

The role of presupposition. The third topic I have considered is presupposition. I have argued that the simple garden variety of negative sentence presupposes that the listener believes in the truth or plausibility of the embedded proposition, but asserts that the listener's supposition is false. A survey of English negatives, however, showed that negatives differed in what they presupposed and what they asserted. For example, John isn't present is a denial of a positive presupposition, whereas John is absent is an affirmation of a negative presupposition; the first is explicitly negative, in the terminology of this paper, whereas the second is implicitly negative. But what is probably more important, there seems to be evidence that Negation Time is consistently different for two expressions that differ only in their assertions and presuppositions (e.g., isn't present and absent). This fact implied that the presuppositions and the assertion of a sentence are an intrinsic part of its semantic representation, and they systematically affect the verification process. Finally, this led me to argue for the scope of negation hypothesis: the greater the scope of negation, the more difficult a sentence is to verify. Although the existing evidence for this hypothesis is very sparse indeed, the evidence does show that a complete model of negation must somehow account for the systematic variation of Negation Time with changes in presupposition.

In short, I have tried to argue that negation is a single unified phenomenon and that there is a single unified theory that accounts for the basic facts about negation. Obviously, however, I have uncovered just as many mysteries as I have solved. Much work lies ahead.

References

Chase, W. G., Young, R., & Clark, H. H. The separation of comprehension and comparison stages in the verification of sentences against pictures. In preparation.

Clark, H. H. Linguistic processes in deductive reasoning. Psychological Review, 1969, 76, 387-404.

Clark, H. H. Semantics and comprehension. In Sebeok, T. A. (Ed.). Current Trends in Linguistics, Vol. 12. The Hague: Mouton, in press.

Clark, H. H. & Chase, W. G. On mental comparisons of sentences and pictures. In preparation.

Clark, H. H. & Offir, C. On 'coming' and 'going'. In preparation.

Clark, H. H. & Young, O. Explicit and implicit negation and its comprehension. In preparation.

Eifermann, R. R. Negation: a linguistic variable. Acta Psychologica, 1961, 18, 258-73.

Gough, P. B. Grammatical transformations and speed of understanding. Journal of Verbal Learning and Verbal Behavior, 1965, 5, 107-111.

Gough, P. B. The verification of sentences: the effects of delay of evidence and sentence length. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 492-496.

Greene, J. M. Syntactic form and semantic function. Quarterly Journal of Experimental Psychology, 1970, 22, 14-27.

Greene, J. M. The semantic function of negatives and passives. British Journal of Psychology, 1970, 61, 17-22.

Gruber, J. S. Studies in lexical relations. M.I.T. Ph.D. dissertation, 1965.

Jones, S. Decoding a deceptive instruction. British Journal of Psychology, 1966, 57, 405-411. (a)

Jones, S. The effect of a negative qualifier in an instruction. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 495-501. (b)

Jones, S. Instructions, self-instructions and performance. Quarterly Journal of Experimental Psychology, 1968, 20, 74-78.

Just, M. & Carpenter, P. Negation and quantification. In preparation.

Klima, E. S. Negation in English. In J. A. Fodor and J. J. Katz (Eds.). The structure of language. Englewood Cliffs, N. J.: Prentice-Hall, 1964, 246-232.

McCawley, J. D. Lexical insertion in a transformational grammar without deep structure. Papers from the Fourth Regional Meeting Chicago Linguistic Society, 1968, 71-80.

Slobin, D. I. Grammatical transformations and sentence comprehension in childhood and adulthood. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 219-227.

Trabasso, T., Rollins, H. & Shaughnessy, E. Storage and verification stages in processing concepts. Cognitive Psychology, in press.

Wason, P. C. The processing of positive and negative information. Quarterly Journal of Experimental Psychology, 1959, 11, 92-107.

Wason, P. C. Response to affirmative and negative binary statements. British Journal of Psychology, 1961, 52, 133-142.

Wason, P. C. The contexts of plausible denial. Journal of Verbal Learning and Verbal Behavior, 1965, 4, 7-11.

Wason, P. C. & Jones, S. Negatives: denotation and connotation. British Journal of Psychology, 1963, 54, 299-307.

Wales, R. J. & Grieve, R. What is so difficult about negation? Perception and Psychophysics, 1969, 6, 327-332.

Young, R. & Chase, W. G. Additive stages in the comparison of sentences and pictures. In preparation.

Footnote

[1] This paper is a distillation of one section out of a chapter called "Semantics and Comprehension" in Sebeok, T. A. (Ed.). Current Trends in Linguistics, Vol. 12, The Hague: Mouton, in press.

Table 1

The Comparison Process (Stage 3) of the "True" Model of Negation

1. Compare the embedded strings of sentence and picture.
    - A. If they match, go to 2.
    - B. If they do not match, go to 1a.

1a. Change value of truth index into its opposite; then go to 2.

2. Compare the embedding strings of sentence and picture.
    - A. If they match, stop.
    - B. If they do not match, go to 2a.

2a. Change value of truth index into its opposite; then stop.

Table 2

Latency Components for the "True" Model of Negation

Sentence Type	Stage 1 Sentence Representation	Stage 2 Picture Representation	Latency Components
True Positive	(A above B)	(A above B)	t_0
False Positive	(B above A)	(A above B)	$t_0 + c$
True Negative	(false (B above A))	(A above B)	$t_0 + c + (b + d)$
False Negative	(false (A above B))	(A above B)	$t_0 + (b + d)$

Table 3

Netation Times for Explicit Full Negation

Source	Example of Negative Sentence Used	Negation Time
Wason	Seven is not an even number.	550
Wason & Jones	Seven is not an even number.	440
Eifermann	Seven is not even (in Hebrew with lo).	736
	Seven is not even (in Hebrew with eyno).	600
Clark & Chase	Star isn't above plus.	685
	Star isn't above line.	709
	Star isn't above plus. (picture first)	556
	Star isn't above plus. (picture first)	660
Clark & Young	The star isn't present.	640
Just & Carpenter	The dots aren't red.	463
Wales & Grieve	Given 6 and 8, the next number is not 1.	455

Median 600

Table 4

Negation Times for Explicit Quantifier Negation

Source	Examples of Negative Sentence Used	Negation Time
Just & Carpenter	None of the dots are red.	306
	There are no red dots.	76
	Hardly any of the dots are red.	385
	Scarcely any of the dots are red.	384
	Few of the dots are red.	210
Trabasso et al.	KOV green.	200
	Median	258

Table 5

Negation Times for Implicit Full Negation

Source	Example of Negative Sentences Used	Negation Time
Clark & Young	The star is absent.	371
	The hole is absent.	185
Chae & Clark	Absent.	145
Clark & Offir	It is coming from town B.	184
	Mary has just come out of the kitchen.	201
	Median	185

Table 6

Negation Times for Implicit Quantifier Negation

Source	Example of Negative Sentence Used	Negation Time
Clark & Chase	Star is below plus.	93
	Star is below line.	135
	Star is below plus. (picture first)	117
	Star is below plus. (picture first)	84
Just & Carpenter	A minority of the dots are red.	241
	A small proportion of the dots are red.	214
	2 out of 16 of the dots are red.	72
Clark & Peterson	Star is lower than plus.	251
Clark	The pink one is in back of the blue one.	189
	Median	135

Table 7

Hypothetical Representations of Sentences and Pictures under Two Conditions

Sentence Type	Sentence Code	Picture Code	Latency Components
True Affirmative	(hole present)	(hole present)	t_0
False Affirmative	(hole present)	(false (hole present))	$t_0 + d$
True Negative	(false (hole present))	(false (hole present))	$t_0 + b$
False Negative	(false (hole present))	(hole present)	$t_0 + b + d$
True Affirmative	(star present)	(star present)	t_0
False Affirmative	(star present)	(plus present)	$t_0 + c$
True Negative	(false (star present))	(plus present)	$t_0 + c + b + d$
False Negative	(false (star present))	(star present)	$t_0 + b + d$

The Psycholinguistic Structure of Knowledge:

1. Some Examples[1]

Edward J. Crothers

University of Colorado

08/01/70

Human memory and cognition involve an integrated network of sentences or logical propositions. Each sentence in turn has its own internal syntactic and semantic structure. Although the structures within a sentence have been intensively analyzed by linguists and psychologists, little attention has been given to the large-scale organization of knowledge; that is, of sentences into larger units. When the issue has been examined at all, the attempts to describe the structure itself have been made chiefly by linguists and investigators in artificial intelligence. Within linguistics, the few efforts to extend the analyses beyond the sentence suffer from an over-dependence on syntax at the expense of semantics. Generative grammarians evidently assume that a paragraph can be transformed syntactically into one long sentence which can then be analyzed just like any other sentence. Any limitations imposed by the added length and complexity of the sentence paragraph are seen as merely technical problems to overcome within the grammar. In fact, this reductionistic assumption is incorrect in principle. Typical paragraphs contain numerous implied assertions that are often essential to the theme. But a grammar cannot represent implications, because it generates only the explicitly stated text. A semantic theory is required. Unfortunately, none are yet sufficiently developed to be applicable to paragraphs. The one linguistic analysis that has actually been seriously applied to paragraphs is Harris'(1952) discourse analysis. The analysis is superficial in that it does not go beyond a segmentation of the text into equivalence classes. It fails to establish a hierarchy of propositions within the text, which can also be attributed to an over-reliance on syntactic principles. Certainly, something like a hierarchy is needed in order to reflect our more or less acute perception of the outline or abstract of a passage. On the other hand, the computer models of human memory begin with a small number of primitive relations, deductive rules, algorithms, and so on and attempt to analyze or synthesize texts (e.g. Jacobson, 1966). To date, this approach has had rather limited scope, largely because many of our normal information processing capabilities have not been explicated and incorporated into computer programs. Recent extensions to paragraph processing (Bobrow, 1968; Quillian, 1969) appear more promising but will not be reviewed here.

In the present paper, the analysis seeks to exhibit all of the relations between sentences in typical descriptive prose. Syntactic and semantic analyses of individual sentences are carried out only to the extent required by the overall analysis. The method of analysis relies heavily on any information which a competent reader might be expected to already have in his semantic memory; it is not limited to syntactic principles or to relations that are readily explicated and incorporated into simulation programs. The goal of the analysis is to provide a formal representation of all the information, explicit or implicit, within the passage. A distinction is made between deep and surface structures, although these terms are not used in their conventional senses. The surface structure includes purely syntactic aspects, e.g. reductions produced by anaphoric forms such as pronouns or demonstratives, as well as sentence conjunctions and other devices whose apparent role is to create a pleasing rhetorical style. The surface structure also includes the linear ordering of the sentences. All of these features are viewed as output phenomena and hence are disregarded in the deep structure. Thus the deep structure is intended to represent the semantic content of the passage. In a modest way, it is also a hypothesis about how a person's memory for the passage is organized. The limitation, of course, is that no analysis is given as to how the text information is embedded in the total semantic memory. The main topic in the present paper is the deep structure. Applications to three prose passages are discussed in detail, and some details have been only tentatively resolved. A full set of general principles cannot be formulated until the present methods have been tested on additional passages. Therefore it seems premature to explore the algebraic properties of the method, although ultimately such analyses may be useful here as they have been with sentence grammars. Also, so far no serious attempt has been made to discover how the traditional rhetorical developments such as comparison and contrast, theme and illustration, etc., are reflected in the structure.

A few remarks about the psychological relevance of this work are in order, although later papers will be devoted more to this matter. Recent psychological studies (e.g. Frase, 1969; Johnson, 1970; Koen, Becker & Young, 1969) indicate that concern with structures in psycholinguistics is beginning to reach the paragraph level. Within the present approach, some of the interesting questions are how understanding and memory depend on the structure, on preknowledge of certain elements of the structure, and also on the correspondence with the surface structure, as well as on numerous task conditions. Memory can be assessed by applying the same analysis to the paragraph produced by a subject when he is asked to recall the passage. To attack these problems, it will probably be necessary to define statistical measures on the structures to be proposed in the present paper, but the selection of suitable measures is an open question. For many purposes, it may be a sounder research strategy to use the proposed methods in order to synthesize paragraphs having desired properties instead of to analyze existing paragraphs.

The Analysis

The analysis will be illustrated by application to three passages, one on oceanography, one on steel production, and one on nebulae.

1. Oceanography Passage

The structure is:

Oceanography is the scientific study of the ocean. Oceanography includes the physics of water and wave movements and the chemistry of water and dissolved substances. It also involves the biology of plant and animal life in the sea, including the vertical and horizontal distribution of sea life. It also includes the geology of the shape of ocean bottoms, and of the layers of earth beneath the ocean bottom. Early research in oceanography consisted of occassional expeditions sponsored by museums or universities. But modern research is a continuous operation by oceanographic research institutes. Oceanographic research is carried out in laboratories, both on the land and on research ships. These ships have special features of construction, such as an inertial guidance system to navigate precisely to a desired area of the ocean. Some ships are powered by nuclear engines, in order to reach remote regions of the ocean. The labs on the ships contain special instruments which can be submerged below the ship. One instrument is a deep-sea camera for photographing the ocean. Another instrument is a current meter for measuring the little understood deep-sea currents. A heat meter is used to measure the heat flow between the land and the sea.

The outcome of the analysis is conveniently depicted as a graph (Fig.2). To aid in understanding it, a simplified approximation appears in Fig.1. The

Fig. 1,2.

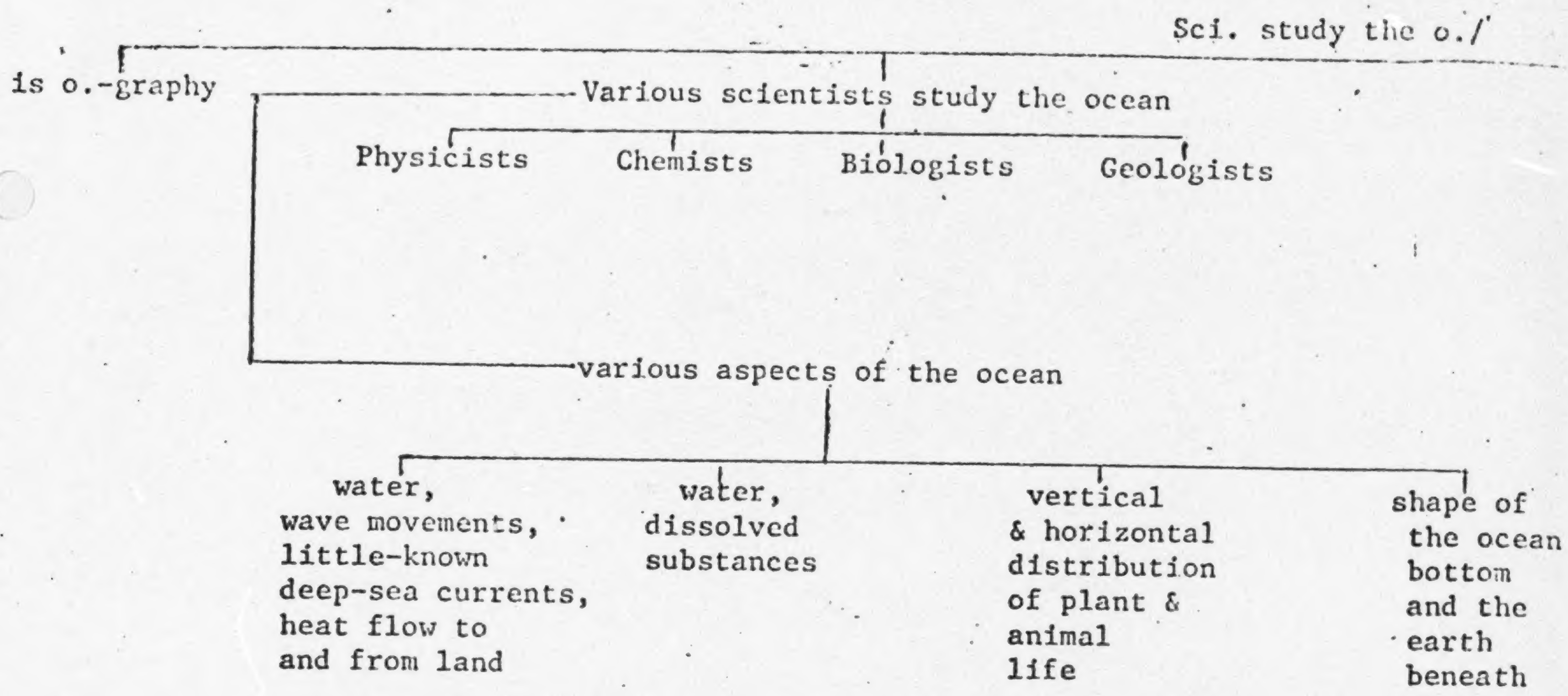

Sci. study the o./
is o.-graphy
Various scientists study the ocean
Physicists
Chemists
Biologists
Geologists
various aspects of the ocean
water, wave movements, little-known deep-sea currents, heat flow to and from land
water, dissolved substances
vertical & horizontal distribution of plant & animal life
shape of the ocean bottom and the earth beneath

Fig.1 Preliminary outline of oceanography passage

Snom, who, when

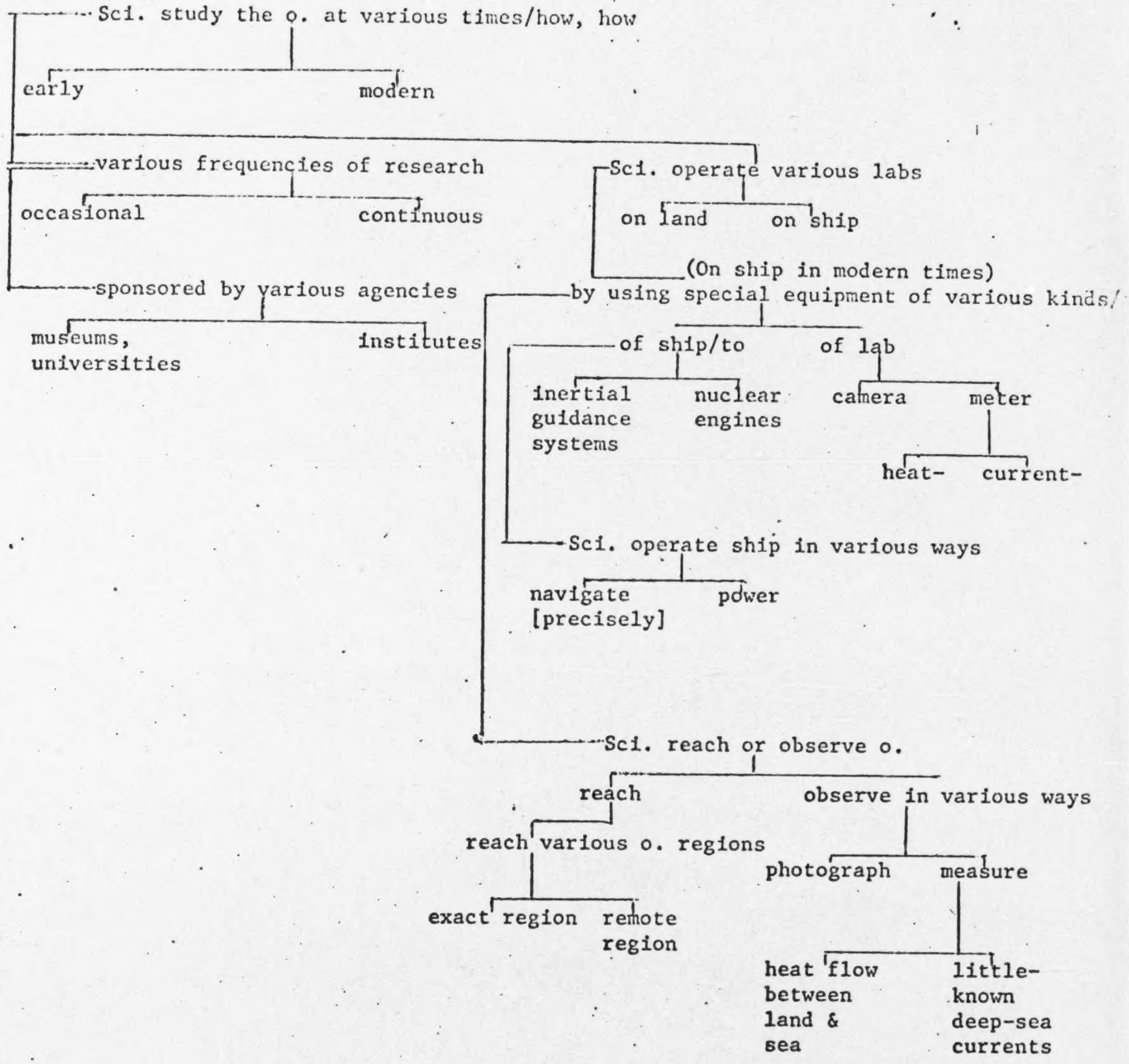
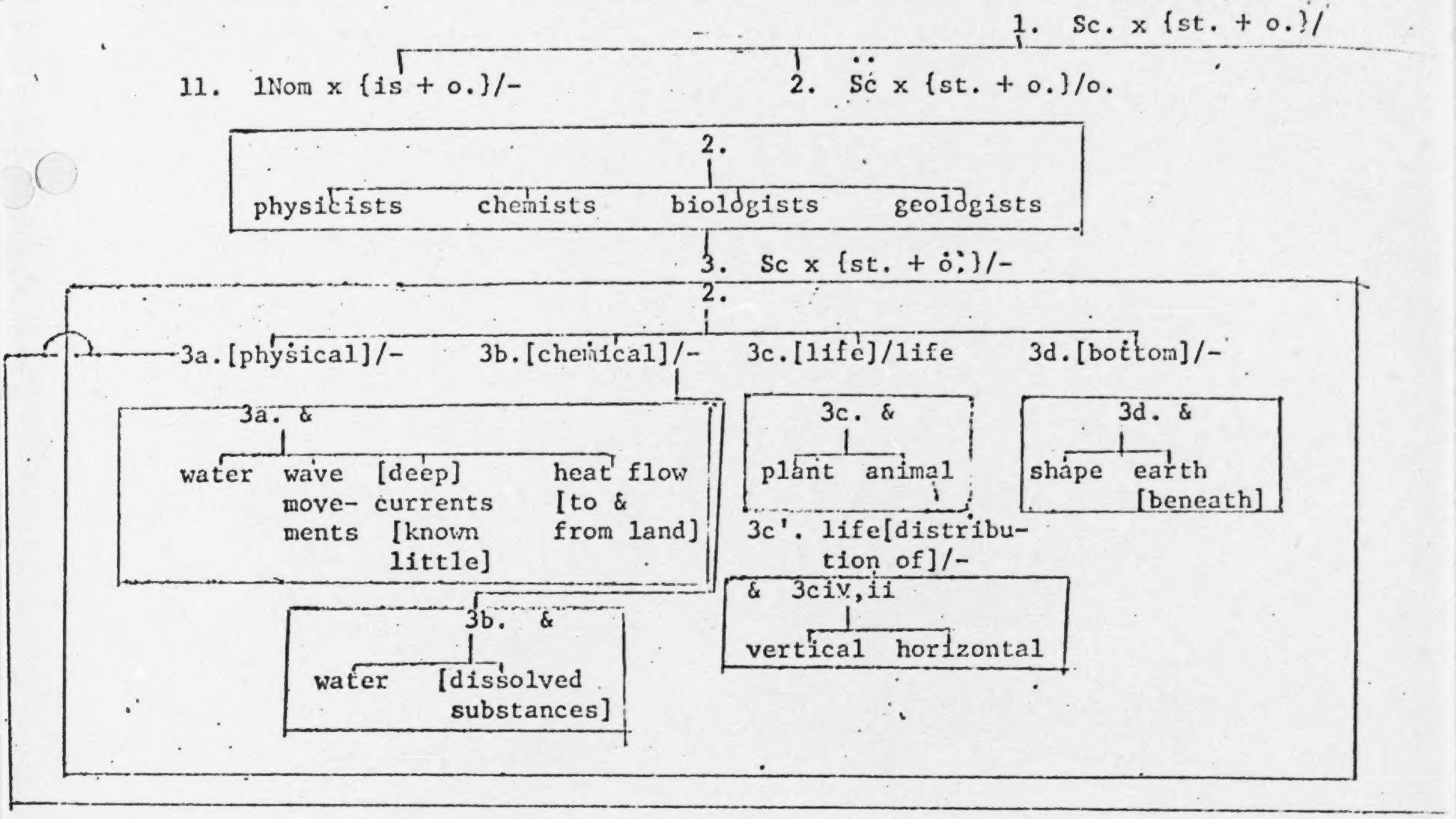


Fig.1 (Cont'd)



Abbreviations

dev.	devices
inst.	instrumentality
o.	ocean, oceanography
Sci.	Scientists
st.	study

Fig.2 Graph of oceanography passage.

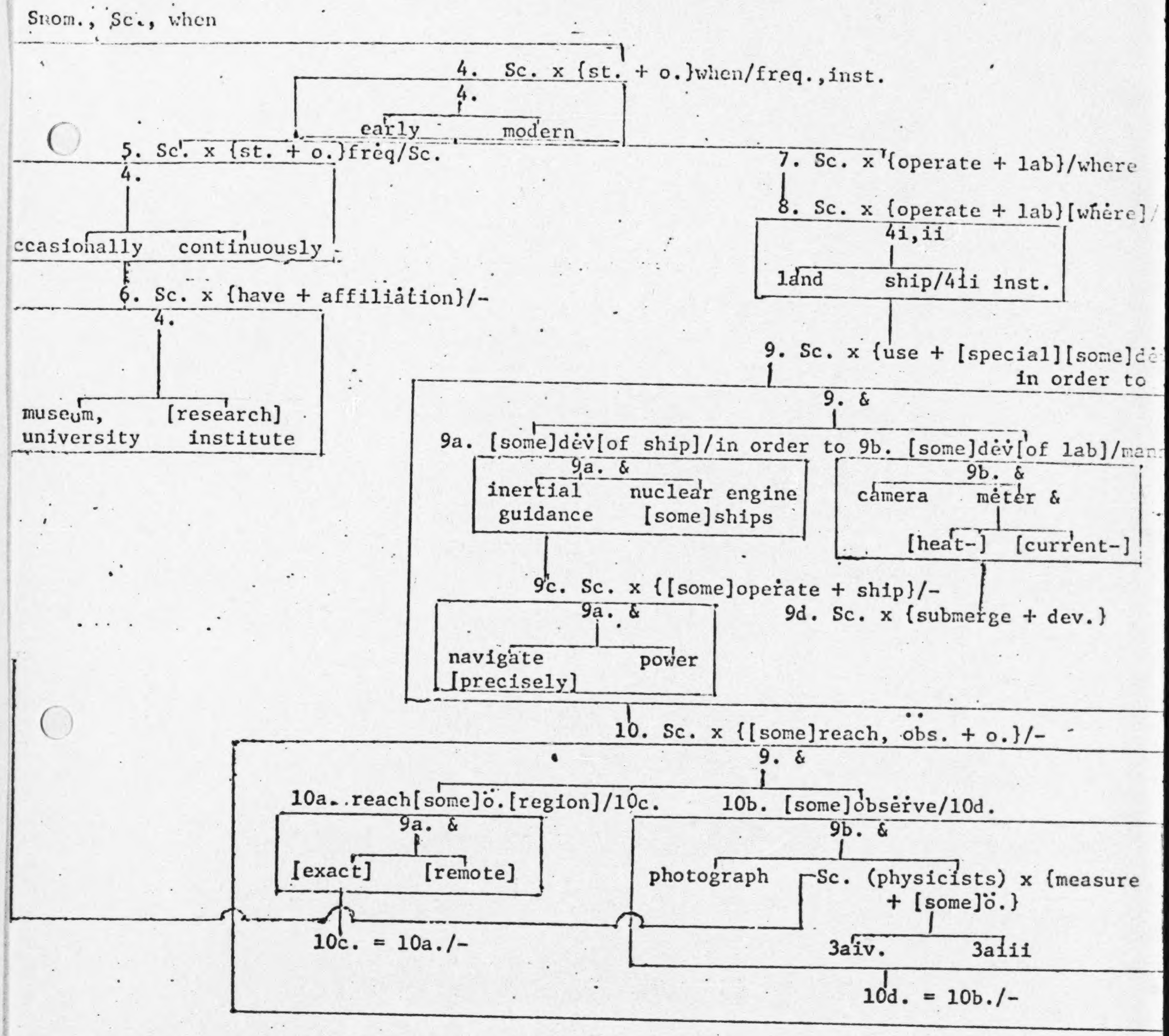
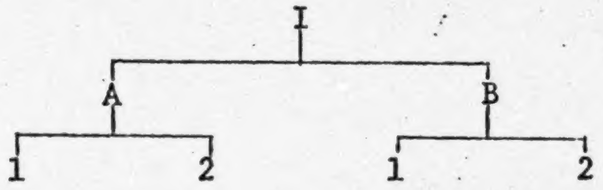
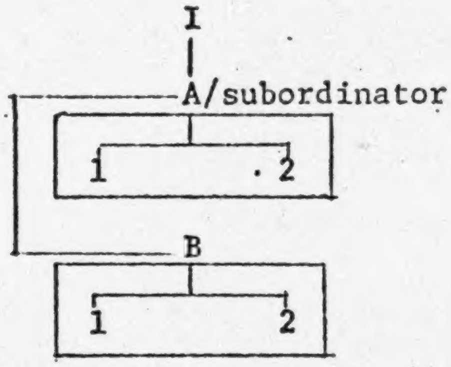

Snom., Sc., when
4. Sc. x {st. + o.}when/freq.,inst.
4.
early
modern
5. Sc. x {st. + o.}freq/Sc.
4.
ccasionally
continuously
6. Sc. x {have + affiliation}/-
4.
museum, university
[research] institute
7. Sc. x {operate + lab}/where
8. Sc. x {operate + lab}[where]/
4i,ii
land
ship/4ii inst.
9. Sc. x {use + [special][some]de
in order to
9. &
9a. [some]dev[of ship]/in order to
9b. [some]dev[of lab]/man
9a. &
inertial guidance
nuclear engine [some]ships
9b. &
camera
meter &
[heat-]
[current-]
9c. Sc. x {[some]operate + ship}/-
9d. Sc. x {submerge + dev.}
9a. &
navigate [precisely]
power
10. Sc. x {[some]reach, obs. + o.}/-
9. &
10a. reach[some]o.[region]/10c.
10b. [some]observe/10d.
9a. &
[exact]
[remote]
9b. &
photograph
Sc. (physicists) x {measure + [some]o.}
3aiv.
3aiii
10c. = 10a./-
10d. = 10b./-

Fig.2 (cont'd)

simplification is an outline, but it differs from an ordinary outline in one important respect. A conventional outline embodies only subordination and coordination, hence yields a tree graph, e.g.



where each node corresponds to a sentence. By contrast, the present analysis is a generalization allowing multiple levels of subordination to any one node, e.g.



If the text does indeed state or imply that B is subordinate to A (e.g. a subordinator such as when or because but not the coordinator and), then the representation of them as coordinates in the first graph is incorrect.

The other important properties are shown in the graphs. Some of the principles will be mentioned now, and the others will be noted as they arise in the detailed comments to follow. Each node contains two kinds of information, the one kind shown as a sentence above a rectangle and the other as a rectangle enclosing a tree graph with labelled nodes. (Some sentences are not followed by boxes). The process is recursive (e.g. within 3 of Fig.2) but for the sake of legibility the more deeply embedded trees are not enclosed in rectangles. The rectangles, in effect are two-dimensional brackets. A rectangle and the enclosed tree will be referred to as a subgraph to distinguish it from the head sentence above it, which will be called an expression. Major expressions are numbered serially, and subgraph nodes are numbered i, ii, etc. from left to right. Thus 4ii in Fig.2 is the node labelled modern. All labelled nodes, both the expressions and the terms within the subgraphs, correspond to sentences, in a way that can be clarified after discussing subordinators. An expression is derived from its immediate superordinate by means of a subordinator, which appears after the slash (/) in the expression. For example, 5 is subordinated to 4 by the word frequency, since the main verb is given in 4 and the adverbial phrase specifying frequency is given in 5. The expression also generates the terms within the subgraph, as follows. Each expression has at most one variable word, denoted by one or two dots above the word. Nodes within the subgraph are generated by assigning values to the variable. By convention, then, variables are introduced successively, one per subgraph, instead of simultaneously (minor exceptions to this are allowed on syntactic grounds). Resuming the question of what sentence corresponds to a given node, two interpretations are possible according to whether the subordinator of the expression is taken into account. If it is not, then the sentence is reconstructed simply by substituting into the expression the exemplar name (value) listed at the node. Thus 8i would be Scientists operate lab on land. On the other hand, by the other interpretation the sentence corresponding to a given node contains not only the present subordinate sentence, but also all superordinates. In sum, the two interpretations are schematized as a. and b., respectively for sentences S1 and S2:

S1 + subordinator	S1 + subordinator
S2	S1 + subordinator + S2
a.	b.

Thus according to b, 8i is the full sentence *scientists study the ocean in early and modern times by operating labs on land and on ships*. While the choice between interpretations a. and b. is not an immediate issue, a logical advantage of b. may be noted. According to it, any sentence (expressions and others also) implies the sentences dominating it in the graph, e.g.

(S1 + in order to) + S2 ⇒ S1

(S1 + relative adverb) + VP ⇒ S1

(S1 + relative pronoun) + embedded NP ⇒ S1.

and even

(S1 + therefore) + S2 ⇒ S1.

(VP = verb phrase, NP = noun phrase)

There are a few apparent exceptions to this principle, mainly involving sentences of the form *NP1 is (named) NP2*.

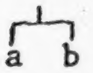
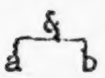
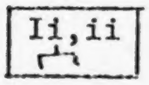
The formal criteria for deciding between subordination and coordination require further study. Basically, subordination seems indicated whenever a sentence or phrase subordinator is present or implied. However, the matter is more complex in that it involves the question of correspondences between subgraphs, to be discussed presently. Once subordination is deemed to be the correct representation, the further question arises as to what is subordinate to what. The direction is not necessarily determined by the subordinator actually stated in the text, because at least with sentence subordinators (as opposed to subordinators such as in S2 = S1 + subordinator + verb phrase) an inverse subordinator can be found (e.g. *S1 in order to S2*, *S2 by S1*). Instead, the order depends on the total paragraph context.

Other general features of the analysis are the following. In some cases an expression is directly dominated by more than one expression. This occurs only once in the oceanography passage, where 10bii is dominated both by 10b and by 3a. However, in other passages it occurs more frequently and cannot readily be replaced by an alternative description. Another important aspect is the pattern of correspondences between subgraphs. Often the text indicates a 1:1 matching of corresponding nodes. For example, node 4i corresponds to 5i in that the *early* research was done *occasionally*. Likewise, 4ii corresponds to 5ii. If the nodes correspond except for a few discrepancies (e.g. no terms or two terms at a node instead of one) the same graph is retained and the discrepancies are marked as will be illustrated in the detailed notes. But if the discrepancies are prevalent, it is more economical to set up a different subgraph. All subgraphs that are headed by the same numeral (e.g. the 4 in both the 4 and 5 subgraphs) indicate the same correspondence. Another type of correspondence is cross-classification, which is defined on subgraphs and not on expressions since the expressions could be related by subordination even if the nodes are cross-classified. In this type of correspondence, each node of one subgraph corresponds to every node of the other, as in 3c and 3c' of the figure.[2]

Other notational symbols are interpreted as follows:

Symbol	Interpretation
•	A variable for which the values in the subgraph are logically exhaustive in the lexical sense, i.e. ignoring the sentence context.
••	A variable for which the values in the subgraph are not lexically exhaustive.

(Continued)

Symbol	Interpretation
[some]NP.	In an expression, the values in the subgraph are not contextually exhaustive, i.e. the subgraph could be made exhaustive by adding a node labelled other.
[all]NP, abbreviated as NP.	The values in the subgraph are contextually exhaustive.
 or a,b	Disjunction of values at different nodes, or the same node, respectively.
 or a & b	Conjunction of values at different nodes, or the same node, respectively.
	The present subgraph is a cross-classification with an earlier one numbered I, for some integer I.
-	Null symbol, at a node or as a subordinator.
[some][ ] or [ ]	Restrictive or nonrestrictive modifier belonging to a different syntactic category than the word modified, e.g. [modern]science.
( )	Modification produced by restriction to a lexical subset, e.g. science(astronomy).
X	Separates main NP and VP of the sentence surface structure.
{ }	Encloses main VP of sentence surface structure.
+	Separates V and substantive (if any) consistituents of main VP of surface structure.
Snom, T	"Transformations" marked when the segmentation into major surface structure constituents is not sufficient for the present purpose.
/S1 ∿ S2 X	X subordinates the clause in the next expression to the entire construction S1[S2], not just to S2.

The symbol [some]NP requires comment. It is usually associated with text phrases such as including, one of, for example, such as, if the complement of the named item is not specified. Sentences 1 and 2 below illustrate the cases.

1. Scientists study ocean heat flow, which is to and from the land.

2. Scientists study ocean heat flow which is to and from the land.

These sentences are marked in the analysis as 1' and 2', respectively.

1'. Scientists study ocean heat flow [which is to and from the land].

2'. Scientists study ocean heat flow [some][which is to and from the land].

For the analysis it is irrelevant that the bracketed clause is always a restrictive modifier of the word flow in isolation. The distinction is between nonrestrictive vs. restrictive modification within the sentence context. Hence the choice between the form illustrated in 1' and the one in 2' depends on the entire sentence. (In the present example 1' happens to be correct).

Notes

Notes are numbered to correspond to the nodes in the figure.

1,11. The main clause in the remainder of the paragraph is scientists study the ocean. Oceanography is only a name for this activity, hence appears subordinated at 11 rather than in 1.

2. A common phenomenon in texts is a statement that declares a variable without giving its values, e.g. There are several kinds of + NP, NP + VP in several ways, NP + VP for several reasons. Two cases arise. If the text eventually enumerates the types, ways, etc. then the prefatory sentence is redundant. In our analysis, such redundant sentences are the ones that contain variables, e.g. expressions 2 and 3 but not the fully specified sentences within the subgraph of 2. The other case, where the text fails to enumerate the instances, is treated merely by marking several, etc. as a quantifier.

A sufficient reason for introducing 2 before 3 is that the partition in 2 is further partitioned in 3.

3c. The 3bi, ii term signifies that the variables in 3b and 3c impose a cross-classification on life.

4. Here there are three variables: the affiliation of the scientists, and the time and frequency of research. To maintain the principle that they be introduced successively rather than concomitantly, some rationale is needed to decide the order of appearance in the graph. Evidently, the time variable should appear first, to reflect the text order. Then come the frequency and affiliation variables in 5 and 6, respectively, adopting the principle that the finer partition (here, into three parts) appear after the coarser one. While there can be no question that 4 dominates 5 and 6, it is not certain that 5 and 6 should not be coordinate.

7. For brevity, the 4 subgraph has been omitted beneath the 7 expression. In this case it is redundant, since the 7 expression would simply be repeated verbatim at both nodes of the subgraph.

8ii. The line from 8ii to 9 indicates that in the remaining text the scope is narrowed to labs on modern ships. Quite obviously the restriction cannot be marked by connecting 4ii (=modern) directly to 9, for the connection between 8 and 9 must not be omitted. The first interpretation is that the author did not impose this restriction earlier, i.e. that 7 and 8 do indeed refer to early as well as modern oceanography. If this reading is incorrect, the alternative would be to connect 7 only to 4ii. This interpretation is explored in the next note. Even if the broader reading is correct, the term land in 8i is debatable.

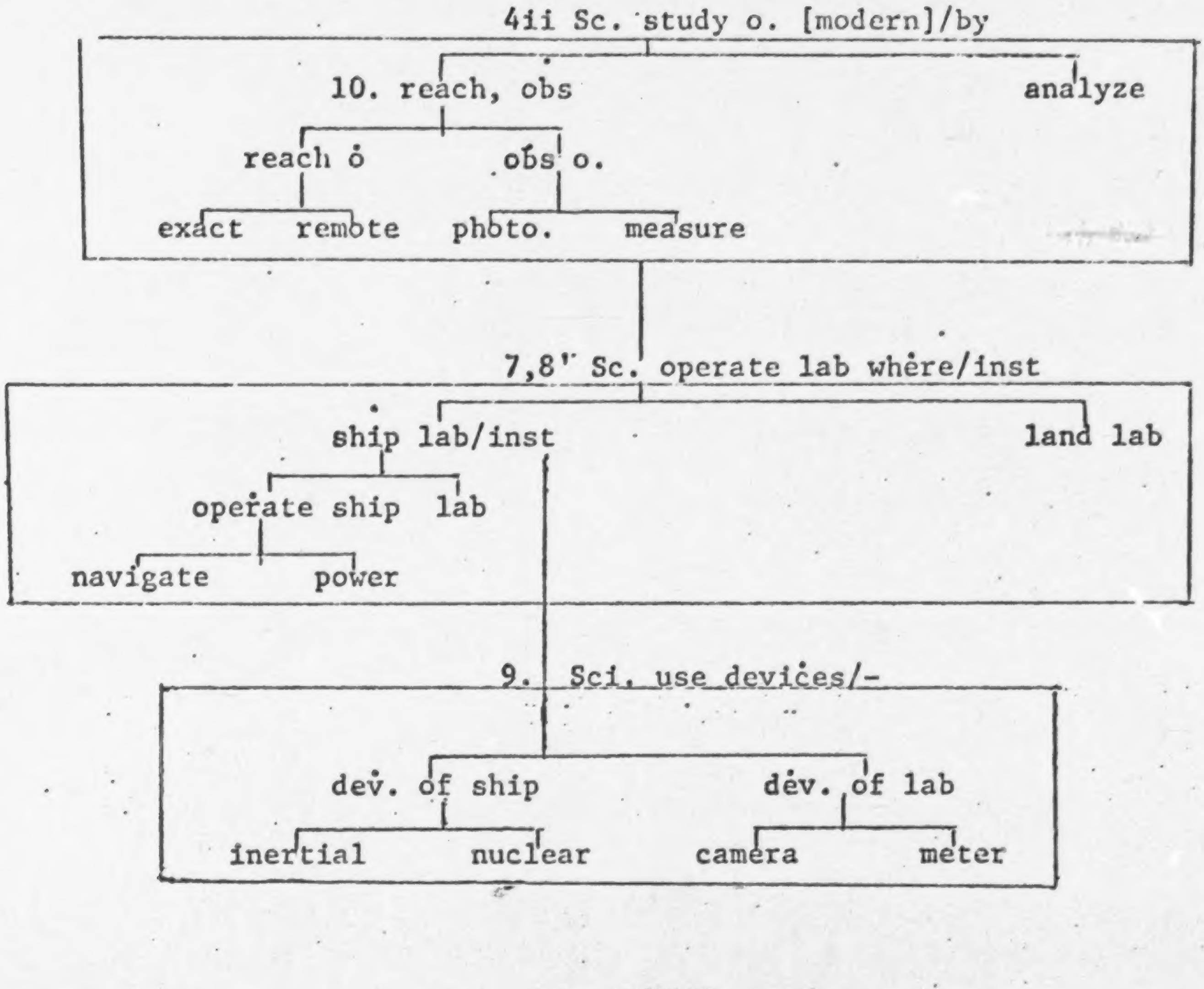
7-11. Here 7,9,10 and 11 each introduces an adverbial clause, and the order of occurrence is an issue. The version in the figure may not be the most insightful one. Instead, a principle might be formulated to the effect that verb phrases are "closer" to each other if they involve the same object NP, as the VP's numbered 4 and 10 do. Hence the revision would be as sketched below:

4. Sci. study ocean/by

10. Sci. reach, observe ocean region.

Now would come 7, 8, and 9.

Of course, permutations of subgraph order entail changes in the adverbial subordinators; here the by replaces freq., inst. Now the inclusion of 7, 8, and 9 becomes cumbersome if parts of these are read broadly as applying to early oceanography also. Hence, assume that all references to labs imply modern research only, namely the alternative suggested at 8ii above. The complete development would be about as follows (omitting the subclasses of measure for simplicity, and collapsing 7 and 8).



and heat flow from 3aiii to 10bii. Of course, one could avoid this type of connection by subordinating all of 4 - 11 to 3. All nodes of 4 - 11 would then occur four times, once for each node of 3, except that the troublesome 10bii would be attached only to the appropriate points of 3a. Such a solution is consistent with the general rationale for claiming subordination, while the method adopted is a modification. Yet the latter might seem clearly preferable, for the fourfold repetition of 4 - 11 would be highly inelegant.

On further thought, a simpler solution would be to exploit the notation already introduced for cross-classification. The subgraphs 4ff. would be headed by the term 3i,iii,iv replacing both the 4 in the 4 subgraph, etc. as well as the line from 3a.

The main syntactic properties of the passage are given in Table 1. For simplicity, all modifiers other than subject modifiers have been regarded as attached to the main verb of the verb phrase. Although this is probably correct for adverbials, it may not concur with modern grammars insofar as sentence subordinators such as because are concerned. On the other hand, even adverbials can serve as sentence subordinators (e.g. when + S, where + S), so it does not seem far wrong to treat because, etc. like relative adverbs.

Table 1 here

The lexical items for Table 1 are:

	Subject N=Noun	V=Verb	O=Object	D=Adverbial
1.	scientists	study	ocean, oceanography	when
2.		operate	lab	instrumentality
3.		submerge	device	where
4.		reach, observe	ship	in order to
5.		have	affiliation	manner
6.			i-iv=physicists, etc.	frequency

The notation is

X → Y	rewrite the most recent occurrence of X as Y.
$\underset{\sim}{\rightarrow}$	same as →, except some details of derivation omitted
*	cross-classification
<a,b> ... <c,d>	correspondence a to c, b to d.
_Z/X → Y	rewrite X as Y in the context Z.
S	Sentence
Xi, Xi(a), Xi(a)(i)	subclass of X, of Xi, of Xi(a), respectively.
Nom	nominalization or quasi-nominal

Table 1

Main Syntactic Properties of Oceanography Passage

Rule	Explanation
1. S → N1 + V1 + 01	Sci. study ocean
11. S → Nom S	Sci. study of ocean
Nom S + be + 01'	Nom S is oceanography
4. V1 → V1D1	when
D1 → D1i, D1ii	early, modern
7. V1 → V1D2S	inst.
S → N1 + V2 + 02	Sci. operate lab
8. V2 → V2D3*D1	where; cross-classification
D3 → D3i,D3ii	land, ship
9. D3 * D1 → D3iiD1ii	Restriction to modern and ship
V2 → V2D2S	inst.
S → N1 + V2 + 03	Sci. use devices
03 → 03i, 03ii	of ship, of lab
03i → 03ia, 03ib	inertial, nuclear
03ii → 03ii(a), 03ii(b)	cameras, meter
03iib → 03ii(b)(i), 03ii(b)(ii)	heat meter, current meter
9c,9d V2 + 03<i,ii> →	gives 9c, 9d
V2 <D4S,D5S> + 03<i,ii>	
S → N1 + V2 + 04	Sci. operate ship
S → N1 + V5 + 03ii	Sci. submerge devices
V2 → V2i, V2ii	navigate, power
03ii → 03ii(a), 03ii(b)	camera, meter
03iib → 03ii(b)(i), 03ii(b)(ii)	heat meter, current meter
10. N1_ 03/V2 → V2D4S	in order to
S → N1 + V4 + 01	Sci.{ reach, observe} o.
V4 → V4i,V4ii	reach, observe
V4i_/01 → 01i, 01ii	exact, remote

	V4ii → V4ii(a), V4ii(b)	photograph, measure
	V4ii(b)_/01 $\underset{\sim}{\rightarrow}$ 07i(c), 07i(d)	heat, current
	V4ii(b)_/N1 $\underset{\sim}{\rightarrow}$ N1S	
	S → N1 + be + 06	
	06 → 06i	Physicists
5.	N1 - 01/V1 → V1D6	frequency
	D6 → D6i,D6ii	occasionally, continuously
6.	_V1D6$^+$01/N1 → N1S	
	S → N1 + V5 + 05	Sci. have affiliation
	05 → 05i,05ii	{mus.,univ.}, institute
	05i → 05i(a),05i(b)	museum,university
2.	_ + V1 + 01/N1 → N1S	
	S → N1 + be + 06	
	06 → 06i - iv	physicists, ... , geologists
3.	01 → 01S	
	S → 01 + has + 07	
	07 → 07i - iv	_, _, life, bottom
	07i → 07i(a) - (d)	water, ... , heat flow
	07ii → 07ii(a), 07ii(b)	
	07iii → 07iii * S	
	S → 07iii + has + 07iii'	life has distribution
	07iii $\underset{\sim}{\rightarrow}$ 07iii(a), 07iii(b)	plant, animal
	07iii' $\underset{\sim}{\rightarrow}$ 07iii'(a), 07iii'(b)	vertical, horizontal
	07iv $\underset{\sim}{\rightarrow}$ 07iv(a), 07iv(b)	shape, earth

Then the overall syntactic structure of a sentence is given by applying all appropriate rules, e.g. for 8i = Sci. operate lab. on land plus the sentences dominating 8i, we have:

1. N1 + V1 + 01

4. N1 + V1D1 + 01

7. N1 + V1D2S + 01

7. N1 + V1D2(N1 + V2 + 02) + 01

8. N1 + V1D2(N1 + V2D3 + 02) + 01

= Sci. + study inst(Sci. + operate on land + lab) + ocean.

A syntactic analysis supplements the graph by identifying both the syntactic basis for the derivation of one subgraph from another, and also the syntax within a subgraph. For psychological processes, it may be important whether the progression is by relative clauses modifying the subject, as opposed to adverbial phrases, or adverbial clauses. Another use is to denote recurrences of particular words within or between subgraphs. Few recurrences occur in this passage, but they are common elsewhere (e.g., in the steel passage).

2. Steel Passage

The surface structure is:

Steel contains iron and carbon. Some forms of steel contain additional elements such as manganese, chromium and nickel. The strength of the steel depends on how the iron and carbon are combined. Specifically, different proportions of iron and carbon are combined by heating and cooling. The strength of the steel depends on the proportion of carbon. If the proportion of carbon is zero the steel is not especially strong. As the proportion of carbon increases the strength of the steel increases. But if the proportion is too high the steel becomes brittle. The strength of the steel also depends on how fast the iron is cooled. Moderately rapid cooling produces stronger steel than slow or very rapid cooling. Moderately rapid cooling produces stronger steel than slow cooling because the particular alloys of iron and carbon that make up the steel depend on the rate of cooling. If the cooling is moderately rapid, strong alloys are produced. But if the cooling is too fast the strong alloys do not have time to form. Or if the cooling is too slow, the strong alloys have time to change to weaker ones. Steel is cooled by quenching the iron. Very rapid cooling is obtained by quenching in water. Moderately rapid cooling is obtained by quenching in oil. Slow cooling is done by quenching in air. The strength of the steel also depends upon how often the iron is heated and cooled. Reheating makes the steel harder and less brittle. Reheating of steel is called tempering.

Figures 3 and 4 give the preliminary outline and structural description, respectively.

Figs. 3 and 4

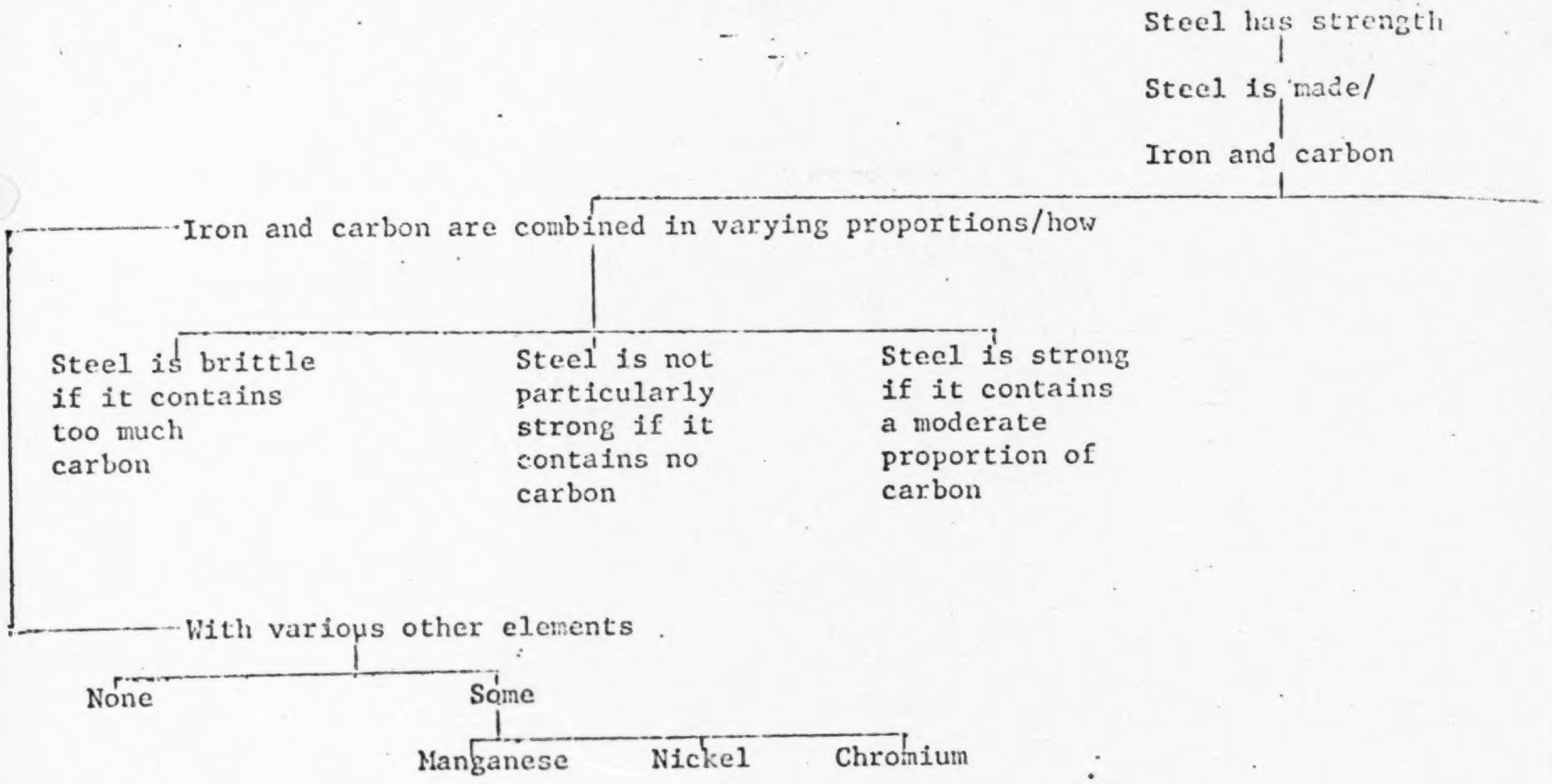
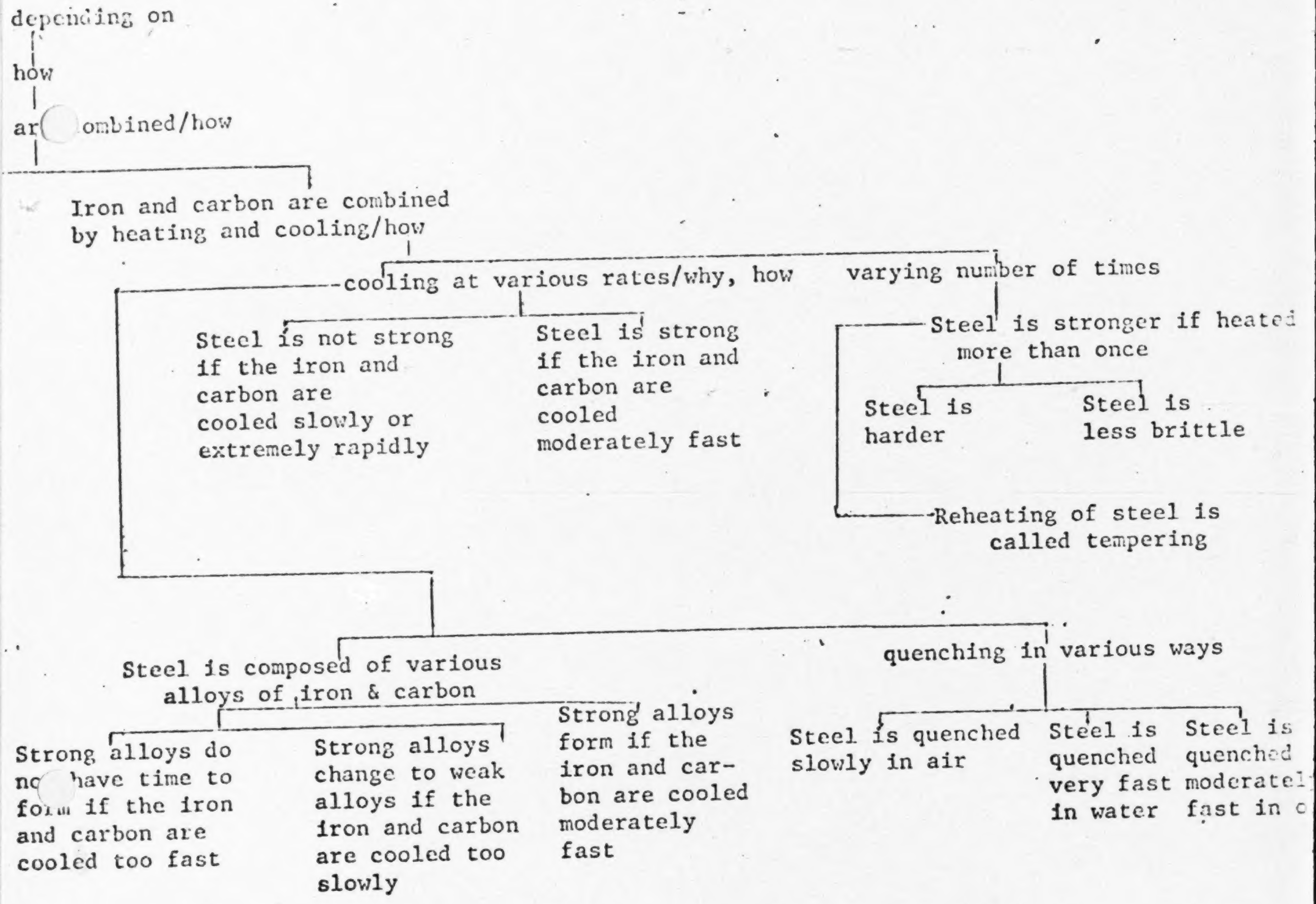

Steel has strength
Steel is made/
Iron and carbon
Iron and carbon are combined in varying proportions/how
Steel is brittle if it contains too much carbon
Steel is not particularly strong if it contains no carbon
Steel is strong if it contains a moderate proportion of carbon
With various other elements
None
Some
Manganese
Nickel
Chromium

Fig.3 Preliminary outline of steel passage



depending on

how

ar ombined/how

Iron and carbon are combined by heating and cooling/how

cooling at various rates/why, how

varying number of times

Steel is not strong if the iron and carbon are cooled slowly or extremely rapidly

Steel is strong if the iron and carbon are cooled moderately fast

Steel is stronger if heated more than once

Steel is harder

Steel is less brittle

Reheating of steel is called tempering

Steel is composed of various alloys of iron & carbon

quenching in various ways

Strong alloys do no have time to form if the iron and carbon are cooled too fast

Strong alloys change to weak alloys if the iron and carbon are cooled too slowly

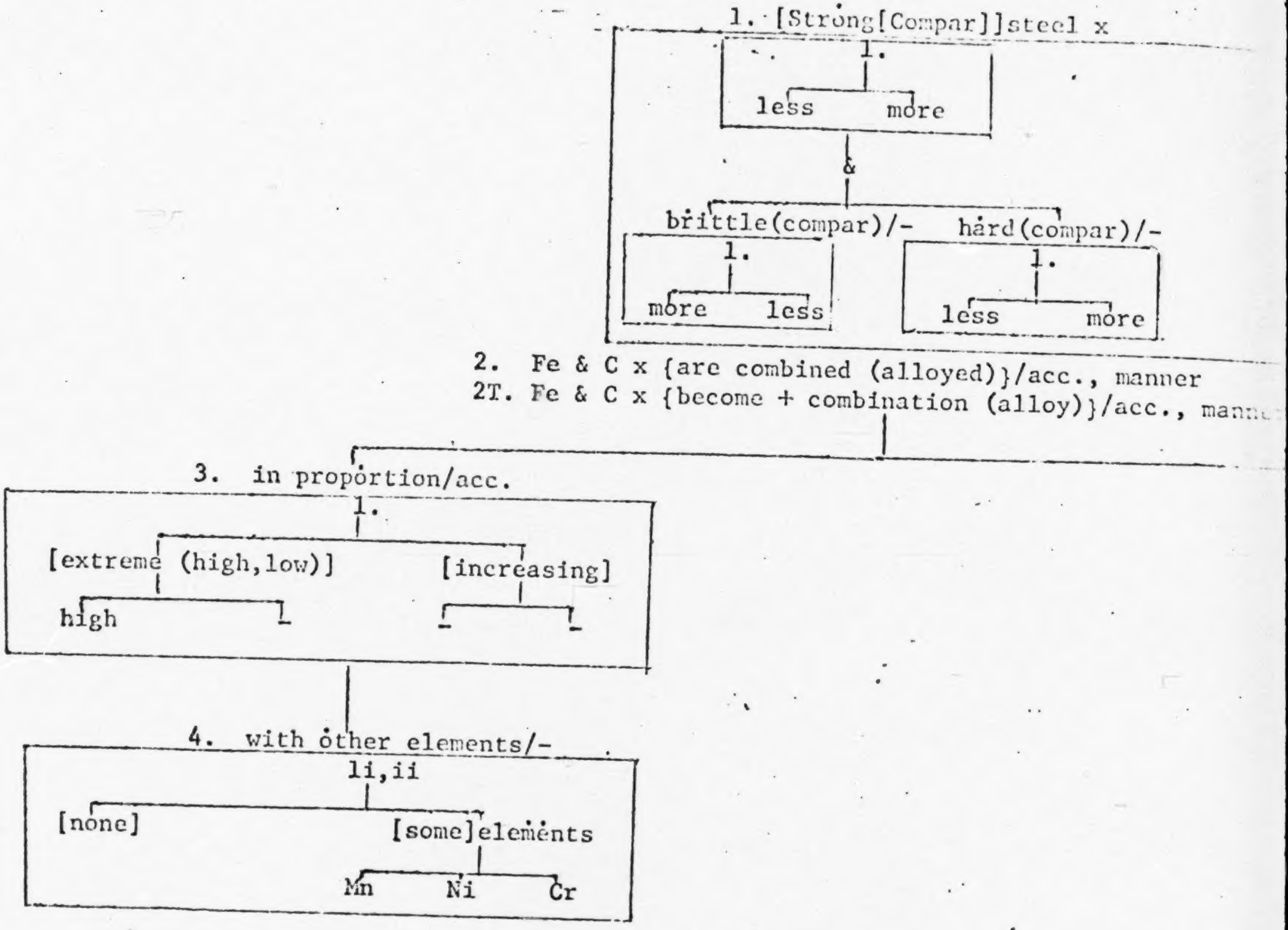
Strong alloys form if the iron and carbon are cooled moderately fast

Steel is quenched slowly in air

Steel is quenched very fast in water

Steel is quenched moderatel fast in c

Fig.3 (Cont'd)



Abbreviations

acc.	accompaniment
C	Carbon
compar.	comparative
Cr	Chromium
Fe	Iron
Mn	Manganese
Ni	Nickel
inst.	instrumentality

Fig.4 Graph of steel passage

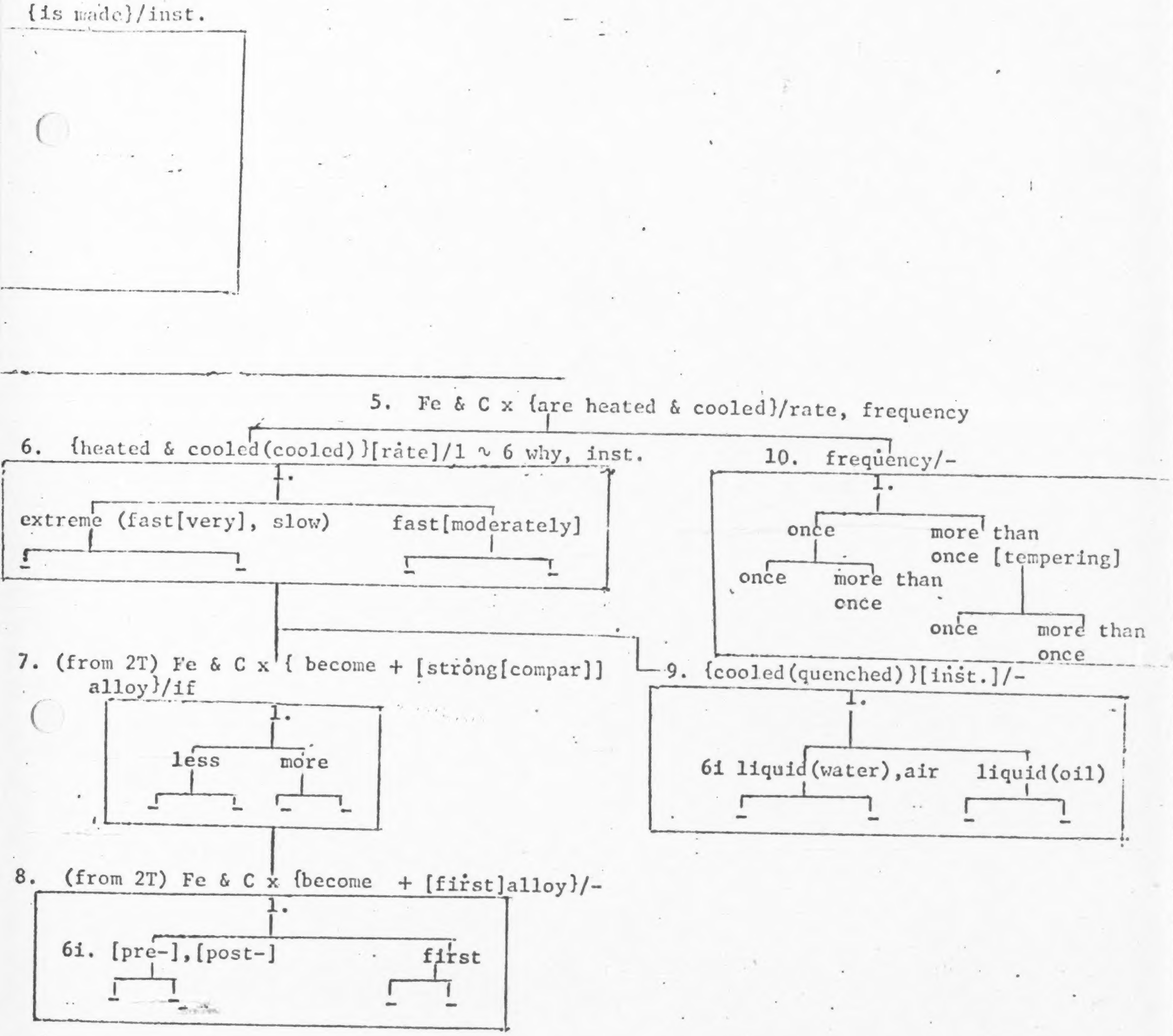

{is made}/inst.
5. Fe & C x {are heated & cooled}/rate, frequency
6. {heated & cooled(cooled)}[rate]/1 ~ 6 why, inst.
1.
extreme (fast[very], slow)
fast[moderately]
10. frequency/-
1.
once
more than once [tempering]
once
more than once
once
more than once
7. (from 2T) Fe & C x { become + [strong[compar]] alloy}/if
1.
less
more
9. {cooled(quenched)}[inst.]/-
1.
6i liquid(water),air
liquid(oil)
8. (from 2T) Fe & C x {become + [first]alloy}/-
1.
6i. [pre-],[post-]
first

Fig.4 (Cont'd)

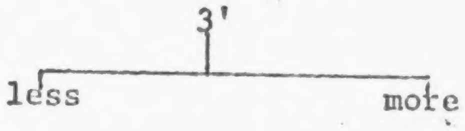
Notes

1. The comparative is chosen because all steel is strong relative to most other materials. The major point to be noted at this node is that the analysis departs from the surface structure by introducing *steel has strength* at the outset. The departure is greatly heightened by introducing the values of *strength* at the same time. Closer adherence to the surface structure incurs complications owing to the subordinate phrases, thus:

1'. S x {is made}/instrumentality

2'. Fe & C x {are combined}/determines

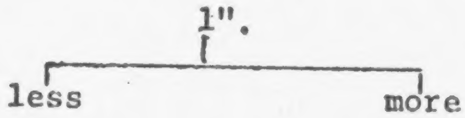
3'. S x {has + strength}/depending on



4'. Fe & C x {are combined}/accompaniment, manner etc. as in the figure.

This version is awkward because of the repetition of the identical sentence at 2' and 4', and also because of the alternation of subject NPs and the anomalous subordinator *determines* at 2'. A more acceptable revision would be:

1". S x {has + strength}/depending on

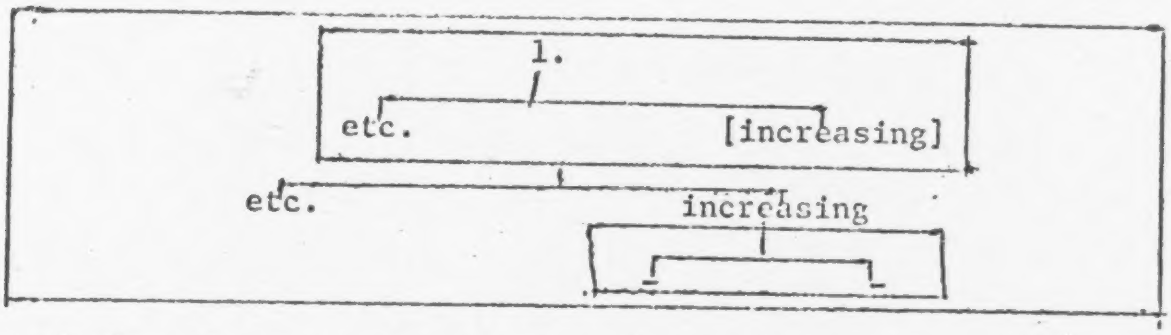


2". S x {is made}/instrumentality

3". Fe & C x {are combined}/accompaniment, manner etc. as in the figure. However, this is essentially the version adopted, differing only in that 1" explicates the derivation of *strong steel* in 1.

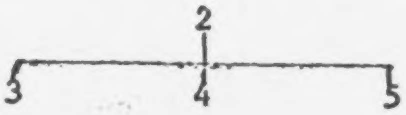
2. The transformation *combined* → *combination* is included so that it can be used at 7 and 8. Also, after the / we continue the policy of not marking the obvious fact that *accompaniment* and *manner* are both adverbials.

3. Details of the correspondence with the subgraph are as follows. The mapping is 1:1 between corresponding nodes of 1 and 3, except that for brevity the expressions *brittle (compar)* and *hard (compar)* have been deleted at 3. The dashes in 3 indicate that no values of *proportion* are specified for these nodes of the *strength* tree. On the other hand, if the text said that increasing the carbon content affected the hardness of the steel, then the 3 subgraph would become

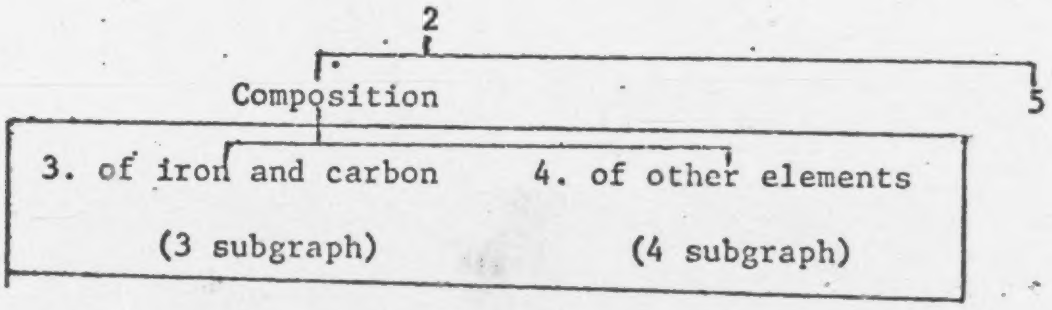


If in particular the text asserted that increasing carbon made the steel more hard, then the lower-righthand blank would be replaced by *increasing*.

3,4. The analysis of 3 and 4 given here is unsatisfactory, since the ordering of the two relative to each other is entirely arbitrary. This suggests that in fact they are coordinate. Of course, the coordination is not simply

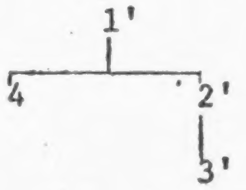


because 3 and 4 both involve the composition of the steel, not the method of producing steel. Hence the proper form seems to be



Several details within the 3 and 4 subgraphs deserve attention. At 3i, both extremes are associated to low strength. But, in the subdivision of 3i, only the extreme high is associated to a subdivision of low strength. The text does not specify how an extremely low proportion of carbon affects brittleness.

The notation 1i, 1ii, in the 4 subgraph indicates a cross-classification of 4 and 1, because the text fails to specify and relation between strength and the presence of these other elements. At first glance, one might be tempted to reinstate the version 1' - 4' which was rejected in note 1. It enables one to avoid the subordination of 4 to strength:



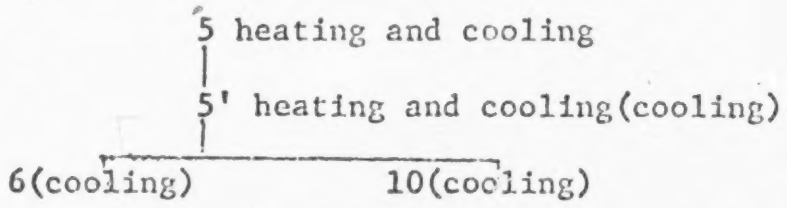
This feature is an advantage because as noted no relationship of strength to 4 is stated. Nevertheless, this alternative still seems inferior. In addition to the reasons given at note 1, the alternative (like the figure, but unlike the improvement in note 3,4) fails to mark the fact that 4 and 3 share the concept composition.

Also, the analysis of 4ii is questionable, because as given it implies a disjunction rather than conjunction of the other elements. Conjunction would be indicated by the term Mn, Ni & Cr at one node. Actually, the text fails to specify whether conjunction or disjunction is intended; perhaps both are.

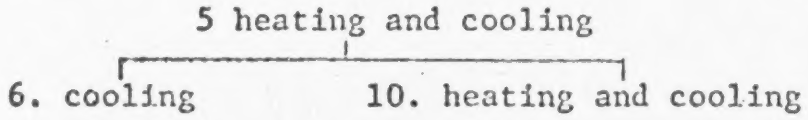
3,6,10. An interesting observation is that the present text gives no information about possible relationships between the proportion of carbon (3), and the rate (6) or frequency (10) of heating and cooling. Thus these variables cannot be in the same path.

5. It would be pointless to replicate the 1 subgraph here, because no values of a variable are introduced.

5,6,7. A very frequent complication in all paragraphs is illustrated here. The restriction from heated & cooled to cooled is definitely required at 6. However, it cannot be made at 5, because heating is involved in making steel. Nor can 5 be retained as is and followed by

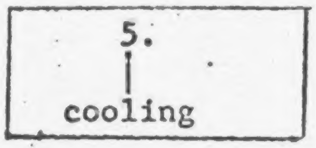


because the restriction to cooling is semantically incorrect at 10. Thus we are left with the awkward but semantically accurate version in the figure, which essentially is



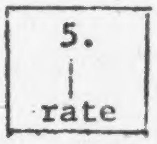
The objection to this analysis is that it scatters the values of a variable, i.e. the elaboration of a particular semantic hierarchy, throughout diverse nodes of the diagram instead of declaring the variable and all values in one single subgraph. Now why cannot a single subgraph be established at 5 and parts of it referred to as needed? In fact the 1 subgraph did exactly that. At 5 the analysis would be

5. heating and cooling

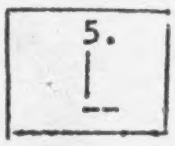


Then this would be cited differently at 6 and at 10, about as follows:

6. rate

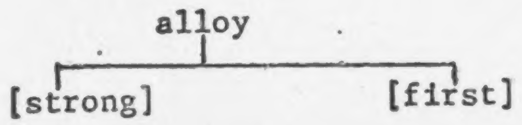


10. frequency



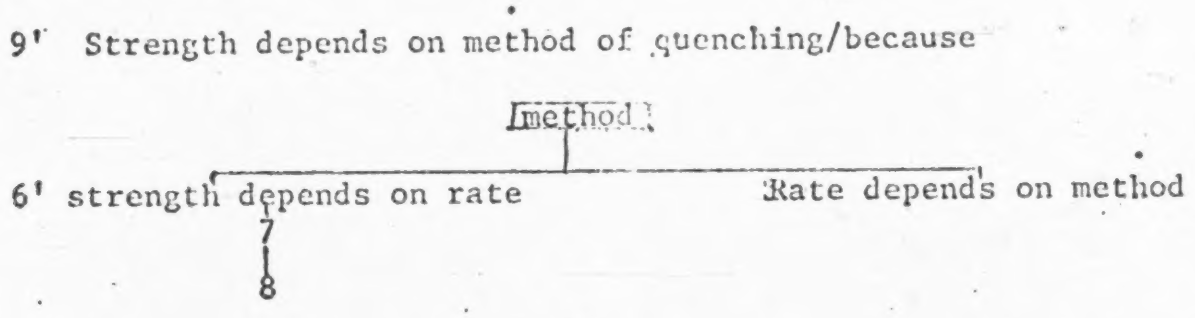
The desired result is achieved by the bottom terms of the 6 and 10 subgraphs, since a correspondence to cooling is given in 6 but not in 10. Unfortunately, the correspondence between 6 and 5 has been achieved at the expense of obliterating the correspondence between 6 and 1, i.e. the 6 subgraph in the figure. In general, a single partition or graph of a new variable can depict its relation to one antecedent variable, but not to several such variables except in the special case where the abstract graph structure is identical for all antecedents. The only solution is to select one subgraph as the template for the others (until it has been restricted, as at 9 in the oceanography passage), and in the present case the frequently recurring subgraph 1 is the logical choice. In addition, there may be some value in enumerating the subgraphs or hierarchies as an adjunct to the main analysis, but so doing would not solve the problem of representing correspondences among more than two hierarchies.

Essentially the same problem often arises at the sentence level. Now it is a question of syntactically related sentences, not related lexical items, appearing throughout the graph. For example, the sentences 7 = Iron and carbon form strong alloy and 8 = Iron and carbon form first alloy. are obviously derived from 2T = Iron and carbon form alloy. The intervention of 5 and 6 obscures the derivational history of 7 and 8. All of the points made above for lexical hierarchies apply when the noun hierarchy is produced by adjectival (or adverbial etc.) modifiers; in this case the implicit tree is



6. The term 1 ~ 6 why after the 1 marks another type of recurrence of earlier sentences. This time the main sentence recurs as part of the object

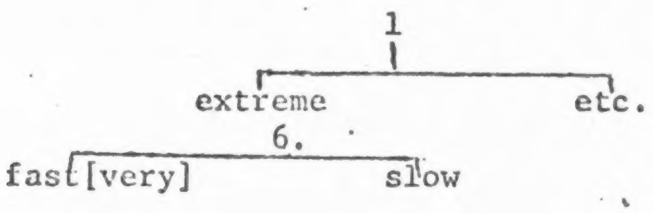
of the interrogative *why*, loosely rendered *Why does the strength of steel depend on the rate of cooling?* Clearly 1 recurs, since the intended query cannot be reduced by eliminating the main clause: *Why are iron and carbon heated at a (certain) rate?* would be a totally different question.
It is important to observe that treating the problem simply by a syntactic distinction between subordinators that attach to a subordinate clause (e.g., most of the ones in this paragraph) vs. subordinators attached to the whole preceding sentence would be a superficial solution. What is needed for psychological purposes is to mark all recurrences of clauses.
6,9. A radically different analysis would be roughly as follows (subgraphs omitted for brevity)


9' Strength depends on method of quenching/because
method
6' strength depends on rate
Rate depends on method
7
8

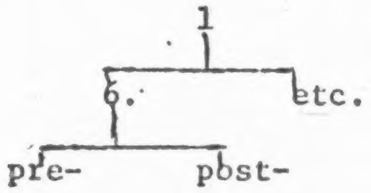
In terms of the overall theme, this differs from the original version by establishing 9 as the principal conclusion, changing both 6 and a new line *Rate depends on method* into subordinate premises. Part of the reason for rejecting this alternative is that making 9 the major conclusion seems contrary to one's intuitive interpretation of the text. The reason is partly formal also, because the *Rate depends on method* brings an extra expression with its subgraph into the analysis. The subgraph would associate *rate* names to nodes of the *method* subgraph, instead of to nodes of 1, the *strength* subgraph. Simplicity dictates that the same subgraph be retained.

7. What is explicit has the form 6 because of 7. A more rigorous logical deduction would add the fact that 7 implies 6. This could be derived from general a priori principles relating properties of wholes to properties of constituents.

8i,9i. The notation 6i signifies a 1:1 correspondence with the two terms of 6i. Again, recourse to such notation is forced because of the decision to make the 6 subgraph correspond to the 1 subgraph. The alternative for 6 would be:



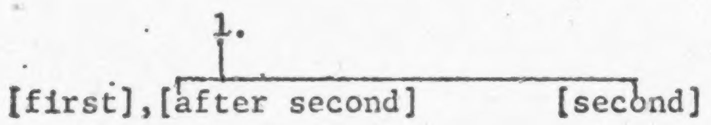
so that 8, and likewise 9, can easily be written



This revision would avoid cumbersome terms such as 6i[pre-],[post]but at the cost of substituting the 6 hierarchy for the lefthand subhierarchy of 1 (the nodes brittle (more) and brittle (less)). Whether or not the revision is an improvement seems uncertain.[3]

Also, the term pre- in 8i is not stated directly, but involves the inference that with too rapid cooling the alloys do not have time to form.

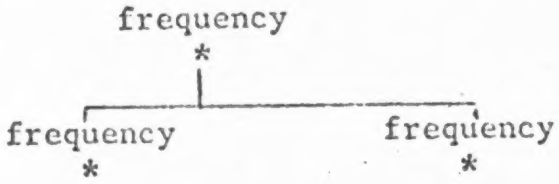
A less plausible but not illogical inference would be that 1. should be relabeled:



Finally, use of the same first term both for the variable and for one value is not ideal.

6,10. There is no conceivable subordinator connecting 6 with 10, so they are coordinate, despite the fact that both assign values to the same partition (viz., the 1 subgraph).

10. Here the policy of abbreviating 1 creates the misleading impression that *more than once* is a value of *once* in 10i, and vice-versa in 10ii. To rectify this, the full form of 1 would give



with the subgraph *once*, *more than once* appearing at each position marked by an asterisk.

A detail omitted is the nominalization *NP + is heated more than once* → *Heating of NP more than once* in order to make it the subject of the sentence whose predicate is *is called tempering*. This correction is needed so that the modifier is not misconstrued as applying to the NP *more times*.

3. Nebula Passage

Two different surface structure versions of the deep structure are possible.

Version 1:

A nebula is any source of light in the sky that has a relatively fixed location in space and looks fuzzy or nebulous. There are two kinds of nebulae. One kind is the nebulae outside of our own galaxy. These nebulae are composed of stars, like our own galaxy is composed of stars. Galaxy nebulae appear in clusters of from two to thirty galaxies. The clusters of galaxy nebulae are spread rather evenly throughout the universe. Galaxy nebulae look fuzzy because the overall nebula is seen, but the nebulae are so remote that their individual stars cannot be distinguished, even with the most powerful telescopes. In fact, with the naked eye only three galaxy nebulae are close enough to be seen at all. The other kind is the nebulae within our own galaxy. These nebulae are clouds of gas or dust. Some of the gas nebulae are evolving to become stars by expanding and contracting. Gas nebulae glow because the gas itself is luminous, but dust nebulae seem to glow because they are illuminated by nearby stars.

Version 2:

A nebulae is any source of light in the sky that has a relatively fixed location in space and looks fuzzy or nebulous. There are two kinds of nebulae, the ones outside our own galaxy and the ones within our galaxy. The nebulae outside our galaxy are composed of stars, like our own galaxy is composed of stars. The nebulae within our galaxy are clouds of gas or dust. Galaxy nebulae appear in clusters of from two to thirty galaxies. The clusters of galaxy nebulae are evolving to become stars, by expanding and contracting. Galaxy nebulae look fuzzy because the overall nebula is seen, but the nebulae are so remote that their individual stars cannot be distinguished, even with the most powerful telescopes. In fact, with the naked eye only three galaxy nebulae are close enough to be seen at all. Gas nebulae glow because the gas itself is luminous, but dust nebulae seem to glow because they are illuminated by nearby stars.

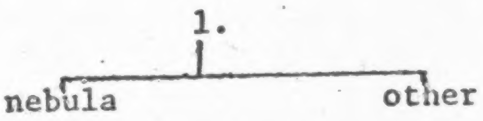
The preliminary outline and graph appear in Figs.. 5 and 6 respectively.

Figures 5 and 6

Notes

1,2 and 9d,10. To make the figure more compact, the path from 1 to 2, and likewise from 9a to 9b, has been drawn horizontally instead of vertically.

1. As in the other passages, the partition of the subject NP is given at the outset. An issue here and elsewhere (e.g., 9b, and 1 of steel) is that sometimes an initial partition gets refined based on a subsequently introduced variable. However, it seems awkward to repeat a previous partition in order to define further partitions on it. Instead, the tentative solution has been to mark the complete tree upon the first occurrence of the variable, thus anticipating that a later justification for the partition will be given. For example, 1 itself only requires the simple subgraph



but the further partition shown at 1 in the figure anticipates the partition to be established at 2.

2. The analysis here is a composite subgraph involving the two variables location and appearence. For the sake of simplicity, the principle of introducing variables one at a time has been violated. Actually, the 2 subgraph is a simplified version of a cross-classification of the trees location and appearence. The simplification was made by collapsing all products except fixed & visible into one term other than fixed and visible.

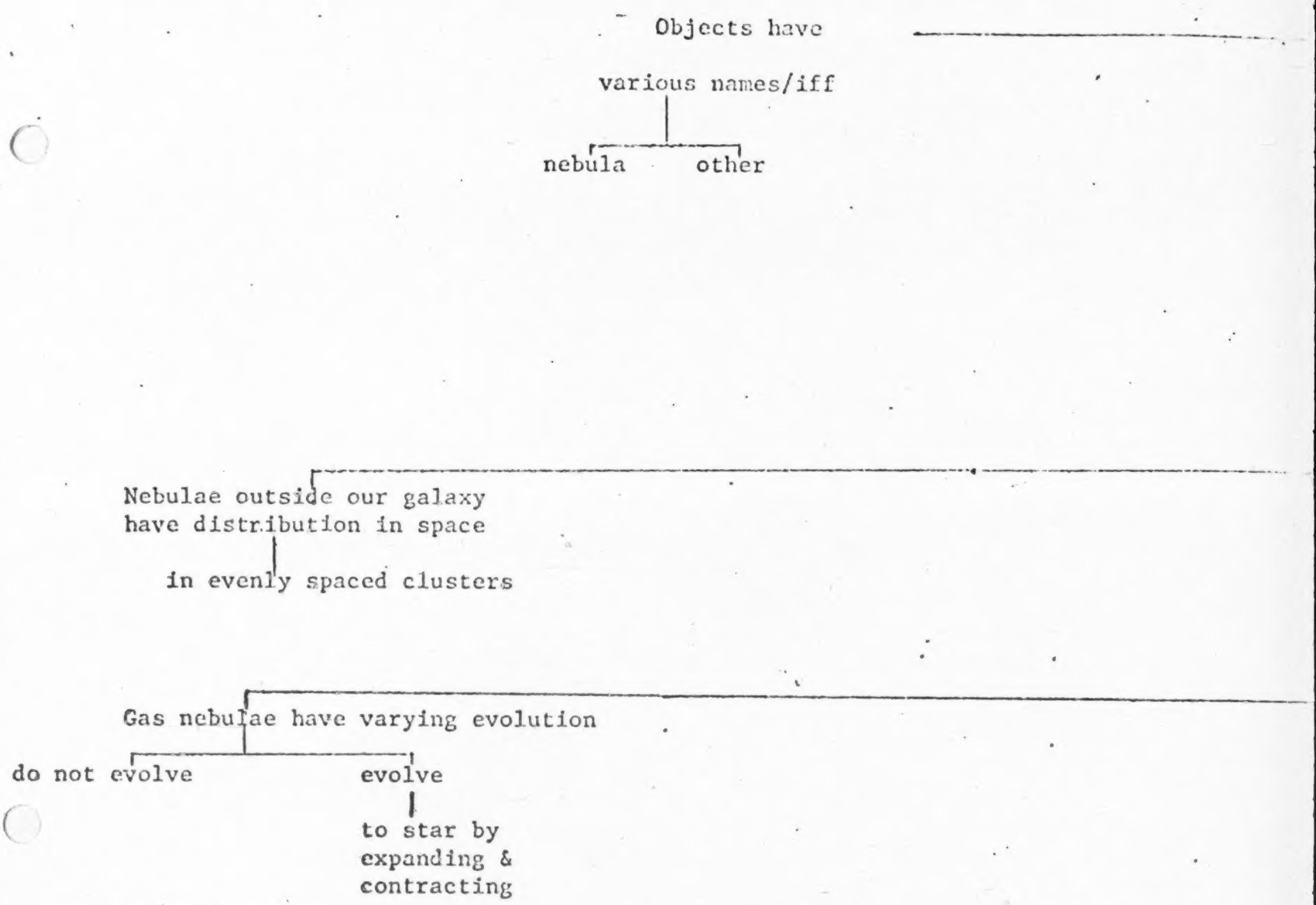

Objects have
various names/iff
nebula
other
Nebulae outside our galaxy
have distribution in space
in evenly spaced clusters
Gas nebulae have varying evolution
do not evolve
evolve
to star by
expanding &
contracting

Fig.5 Preliminary outline of nebula passage

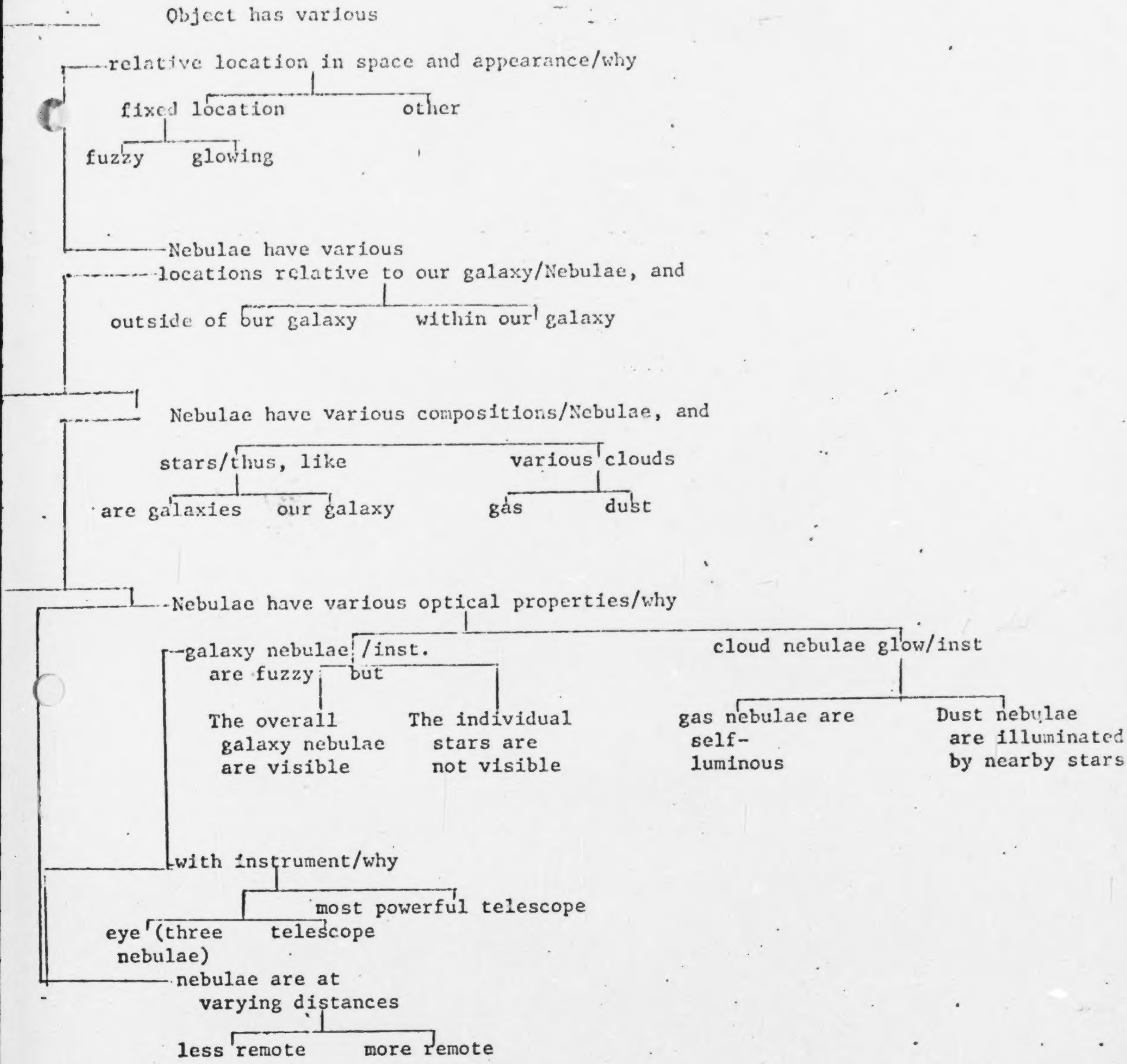

Object has various
relative location in space and appearance/why
fixed location
other
fuzzy
glowing
Nebulae have various
locations relative to our galaxy/Nebulae, and
outside of our galaxy
within our galaxy
Nebulae have various compositions/Nebulae, and
stars/thus, like
various clouds
are galaxies
our galaxy
gas
dust
Nebulae have various optical properties/why
galaxy nebulae /inst.
are fuzzy but
The overall galaxy nebulae are visible
The individual stars are not visible
cloud nebulae glow/inst
gas nebulae are self-luminous
Dust nebulae are illuminated by nearby stars
with instrument/why
most powerful telescope
eye (three nebulae)
telescope
nebulae are at varying distances
less remote
more remote

Fig.5 (Cont'd)

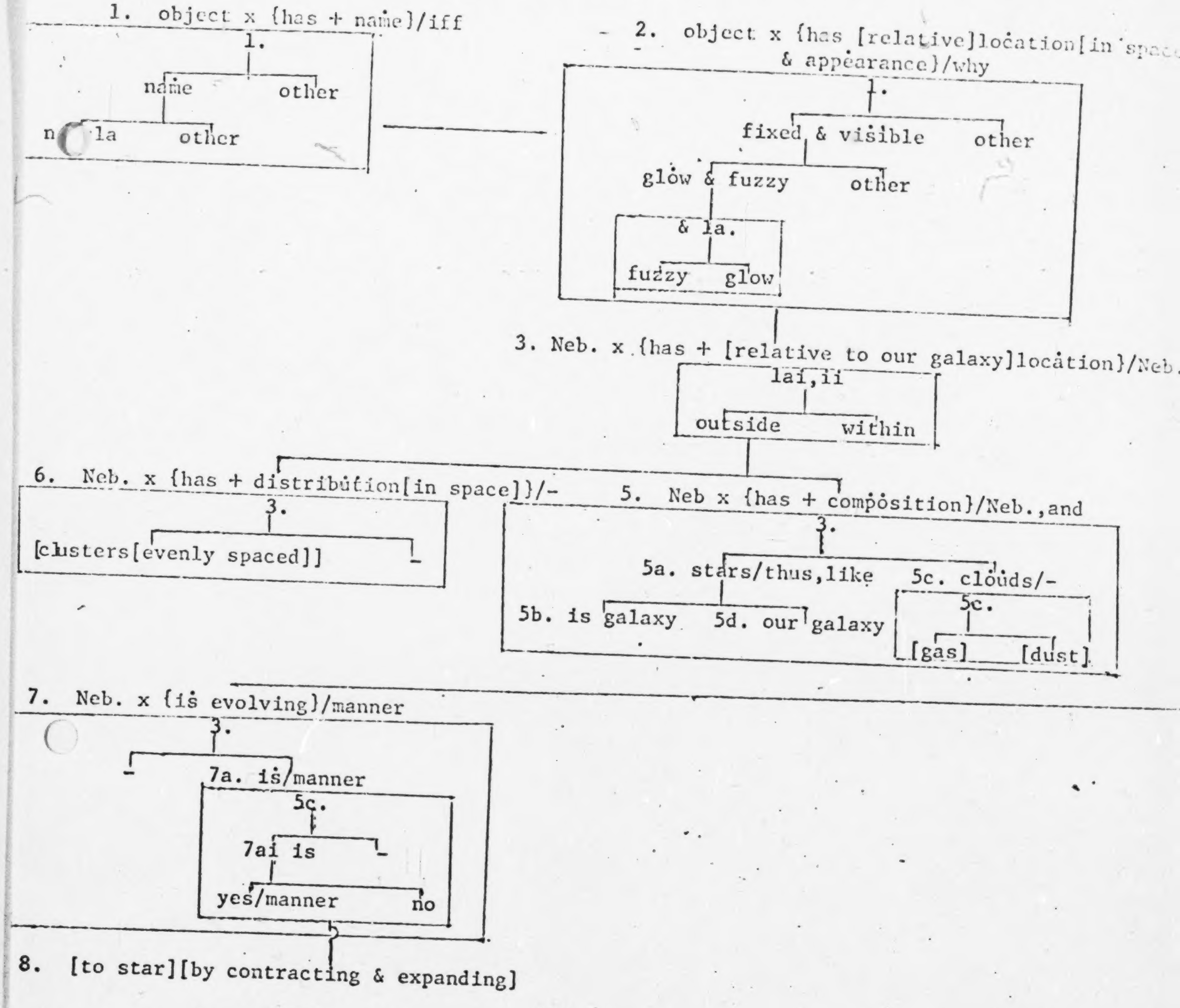
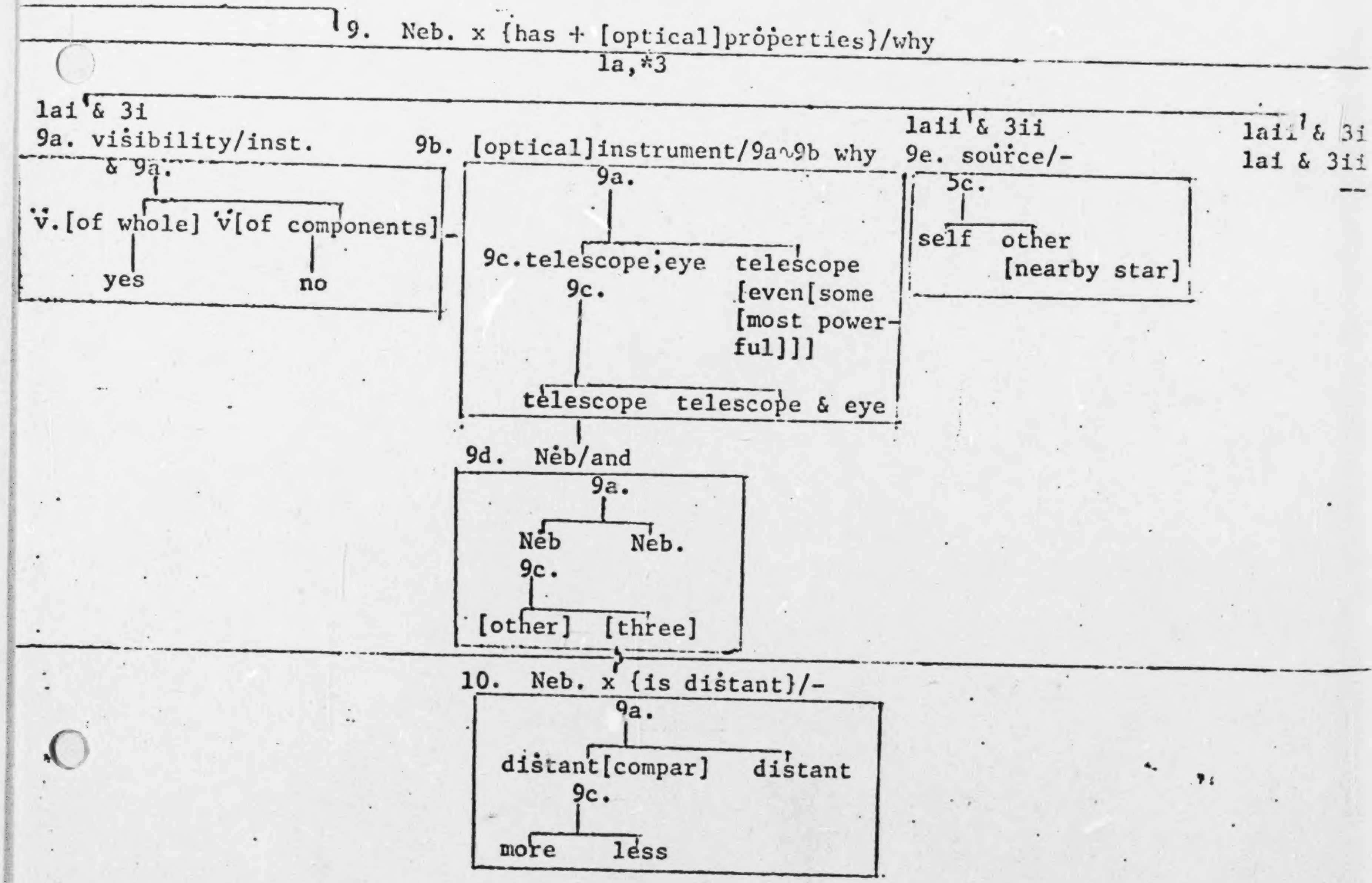

1. object x {has + name}/iff
1.
name
other
n la
other
2. object x {has [relative]location[in spac
& appearance}/why
1.
fixed & visible
other
glow & fuzzy
other
& 1a.
fuzzy
glow
3. Neb. x {has + [relative to our galaxy]location}/Neb.
1ai,ii
outside
within
6. Neb. x {has + distribution[in space]}/-
3.
[clusters[evenly spaced]]
-
5. Neb x {has + composition}/Neb.,and
3.
5a. stars/thus,like
5c. clouds/-
5b. is galaxy
5d. our galaxy
5c.
[gas]
[dust]
7. Neb. x {is evolving}/manner
3.
-
7a. is/manner
5c.
7ai is
-
yes/manner
no
8. [to star][by contracting & expanding]

Fig.6 Graph of nebula passage

Fig.6 (Cont'd)

Abbreviations

compar.	comparative
inst.	instrumentality
Neb.	nebula(e)
v.	visibility



A second point is that in 1a the terms glow and fuzzy are semantically interpreted as being joined by conjunction rather than implication. Some objects glow and are not fuzzy, and conversely.

Another issue is how to mark the fact that 1 and 2 are true by the definition of a nebula. Tentatively, this has been done by assigning the label other to all remaining nodes of 1 and 2, effectively establishing an if and only if convention between nebula in 1 and glow and fuzzy in 2. Definitions apparently do have the logical form Definiendum iff definiens. Note also that the subject NP in 2 is object, not nebula. The latter would be incorrect, for then the answer to why would not be 3ff., but simply by definition.

3. Nodes 3ff answer the question why(=because of) in the 2 expression. Again, the logical form is: To prove :B; Proof: A and A $\Rightarrow$B. What is to be noticed is that the graphs list A explicitly, but not the statement A $\Rightarrow$B.

A principle characteristic of this passage is the manner in which nodes 3ff. answer the why of node 2. The details here are somewhat complex and difficult to represent, although the present version seems generally satisfactory. In order to clarify what the problem is, it will be helpful to think of nodes 3(location) and 5(composition) as one subset, nodes 6(distribution) and 7(evolution) as a second subset, and nodes 9ff. as a third subset. (There is no node numbered 4). On first glance, it appears that only nodes 9ff. actually answer the question of why the nebula glow and are fuzzy, while both 3,5 and 6,7 introduce other properties of nebula. But a more careful reading indicates that only nodes 6 and 7, not 3 and 5, are irrelevant to the why of 2. Nowhere does the text imply that the spatial distribution or evolution of the nebula are at all responsible for the glowing, fuzzy appearance. However, the facts that nebulae are composed of stars or clouds (of gas or dust) is indeed part of the reason why nebulae glow and are fuzzy. Thus 5 belongs subordinated to the why of 2. Moreover, the tree given at 5 was first established at 3(=nebulae outside, within our galaxy), so 3 also answers the why of 2.

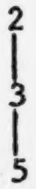
To amplify this final point, the paradigm is

2. B/why

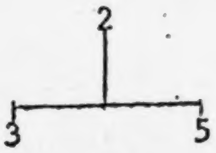
3. $B = B1 \cup B2$

5 and 9ff. $A = A1 \cup A2$ and $A1 \neq A2$

Along with the implied expressions $A1 \Rightarrow B1$ and $A2 \Rightarrow B2$. The point is that line 3 is regarded as a node in the path of inference, i.e. (a) instead of (b),



(a)

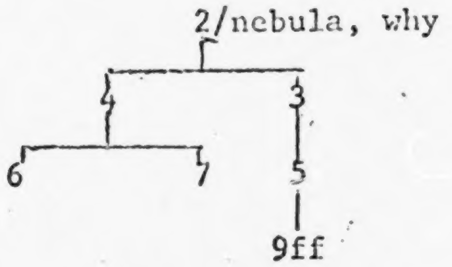


(b)

where (b) invokes the left-hand side of the $A = A1 \cup A2$ expression. The conclusion, then, is that the functions of nodes 3ff. are

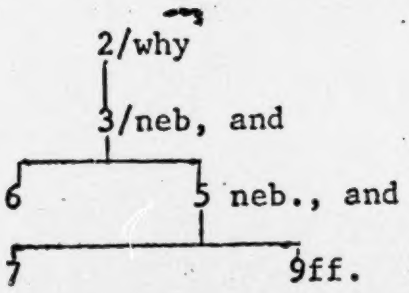
nodes	answer why of 2	give properties of nebula
3,5	+	+
6,7	-	+
9ff	+	+

How can the conditions in this table be represented graphically? Evidently the properties column is of little importance, since physical properties is a very broad semantic notion and only discriminates 1a from the undifferentiated set 3ff. One could depict this at once by enclosing all of 3ff. in a subgraph, but this minor modification has not been made. The major issue is the why column of the table. One attempt to represent it can be schematized as



where some node 4 is needed because 6 and 7 are not related to each other by subordination. This solution is either incorrect or incomplete, because 6(distribution) is undeniably derived from 3(classification by location), and also 7(evolution) is clearly derived from 5(composition). One way to salvage the solution, then, might be to add the connections from 3 to 6 and from 5 to 7. However, the improvisation is questionable, both because it fails to mark the direction (unless a convention about node numbers is added) and because the inserted node 4 seems meaningless. All that 6 and 7 have in common is the negative aspect of not answering the why of 2.

The solution adopted in the figure shifts more of the burden to the interpretation of the words used as subordinators. Schematically, it is



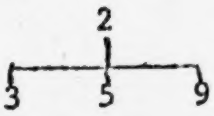
The critical terms are the pair of subordinators neb., and that appear both at 3 and at 5. Here and is treated as having a special role, namely to add a further statement toward answering the why of 2. On the other hand, subordinators other than and (here, neb.) signify that the why of 2 is no longer being answered. The and essentially propagates the why of 2 to the lower nodes. Thus we might consider replacing and with why. However, that suggestion seems misplaced here because a sequence of why subordinators has been reserved as the notation for the case where each why is answered by the next subordinate, i.e. A because of B, B because of C, C because of D = $A \Leftarrow B \Leftarrow C \Leftarrow D$, etc. Such an inferential chain is clearly not suited to the present paragraph, since in

$2 \Leftarrow 3 \Leftarrow 5 \Leftarrow 9$

the second and third implications grossly misrepresent the semantic character of the passage. A much better paradigm is

$2 \Leftarrow 3 \,\&\, 5 \,\&\, 9$,

hence the adoption of *and* subordinators. Yet the conjunction cannot be graphed as



because 3 states the basis for the partition in 5 and 9 and 5 refines the partition in 3. A second tactic aimed at reserving the *and* for coordination would be to rewrite the paradigm as

$2 \Leftarrow 9 \Leftarrow 5.\ (\Leftarrow 3?)$

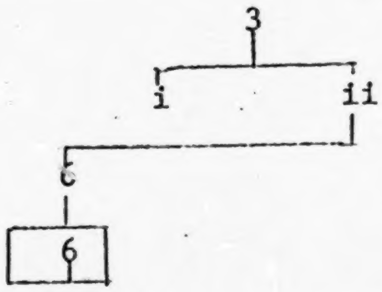
A sufficient reason for rejecting this is, again, that 9 relies on the partition already established at 3 and 5.

Incidentally, note that 3 is indeed a cross-classification with 1, because the text indicates that *fixed*, *glow*, *fuzzy* apply both to nebulae within and nebulae outside our galaxy.

3,5. Again, the rationale for placing 3 before 5 is that 5 subclassifies 3, and perhaps also the fact that the text progression is e.g. *Nebulae which are outside our galaxy are composed of stars*, not *Nebula which are composed of stars are outside our galaxy*.

5a. It is perhaps unnecessary to set off as expressions any terminal predicates which apply to only one node. Instead, all such constructions regardless of syntactic properties could be written as parenthetic remarks and enclosed in [ ], as is done at 9bi,ii and elsewhere. On the other hand, another criterion would be to set off the expression whenever it has a further nontrivial syntactic analysis. Thus 6 and 8 would be set off as shown, and also 5b.

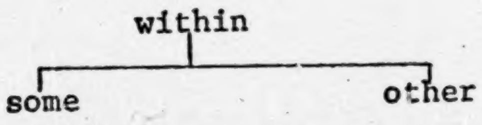
6. An equally acceptable alternative would be, as in 3



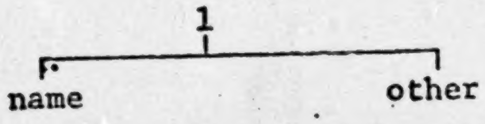
This style is more economical when the restriction discards many nodes, but when only one is discarded it can be identified by a null symbol (-) at the terminal as in subgraphs 6,7,7a, and elsewhere.

5c. A secondary point implied here is that our galaxy is not regarded as a nebula itself. This point is consistent with 2, because our galaxy does not fulfill the conditions given there.

5c. The tree 5c is a partition of 3. The analysis might appear inconsistent in that this partition was not anticipated at 3, unlike the method in 1 and 9b. On further thought, there is really no inconsistency, because 3 is a different case than 1 or 9b. That is, the "anticipatory" version of 3ii would be



The terms some, other can be suppressed as in the figure, because they are immediately recovered once 5c is stated (as a disjunction). By contrast, if 1 were reduced to



then the term *nebula* could not be retrieved. Instead, it would be replaced upon expansion by *some*. A similar situation arises at 9b.

7. *Nebula* seems preferable to *composition* as the subject of the expression, both on semantic grounds and because the former permits reference to the 3 tree.

The graph reflects the statement that *some* gas nebulae evolve, together with the absence of any information about possible evolution of other types of nebulae.

8. A small detail is that the recurrence of *star* from 5a is not marked. This does not seem serious, since the semantic connection between *evolves into a star* and *is composed of stars* seems rather remote.

9. The 1a * 3 tree could have been introduced earlier, at 3,5,6, and 7, [illegible] deferring it avoids the repetition of the same labels at different [illegible]. The reason why the nodes 1ai & 3ii and 1aii & 3i are null is that [illegible] text does not say why star nebulae glow, or why cloud nebulae appear [illegible] respectively. An important issue arises here, and it can be put [illegible] for the sentence *star nebulae glow*. The reason why this is [illegible] in the definition of a star, hence is common knowledge. [illegible] knowledge implied in the paragraph? Only if an affirmative [illegible] does the knowledge meet our criterion for inclusion in the graph.

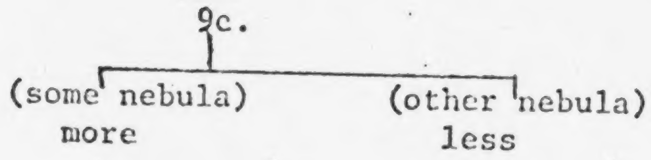
Evidently the question boils down to whether implied semantic information consists only of the inferences logically deducible from the text, or if in addition it includes definitions and other propositions which an average reader can be expected to know. The question remains open. The version tentatively adopted excludes such external lexical information, i.e. as star glows/why:_ instead of star glows/why: star is source of light. Cf. the second argument in 9e.

9a. The prepositional phrases of whole - yes and of composition - no express the statements The overall nebula is seen and The individual stars of the nebula are not seen, respectively. An alternative syntax would have been to replace the 9 expression by Neb, composition of neb. + VP. However, this adds another variable beyond the one already in the VP, hence seems less desirable. There is another matter that neither version handles adequately, namely the fact that composition is a recurrence from 5. Perhaps a solution would be to expand 9aii = v[of composition] as

9aii
|
5. no

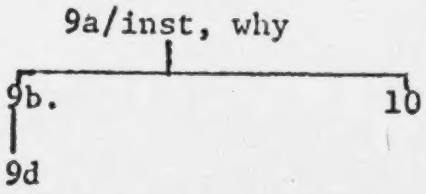
Observe that the lower nodes of the 5 tree need not be written, since they would all be labeled no.

9b,9d. The completion of 9b is rendered somewhat awkward by the insertion of 9d, which is like 3 in partitioning the nebula preparatory to answering the why. However, unlike the partition in 3, the one in 9d does not add much new semantic content and can perhaps be omitted. Then 10 would follow at once after 9b, and the subgraph 9c of 10 already implies some, other:



What has been lost by deleting 9d is the specification of some and other as three and other than three, respectively. Unfortunately, there is no convenient way to reinstate that information at 10 (under the convention of only one variable per subgraph) so perhaps the 9d subgraph is indeed needed in the graph.

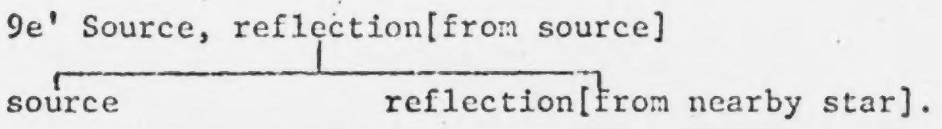
9b. The rather cumbersome notation 9a ∿ 9b why indicates that the question why refers to both the main clause (9a) and the subordinate clause (9b), i.e. the visibility as well as the choice of viewing instrument are interrogated. A similar case arose at node 6 of the steel passage. On the one hand the subordinator cannot be reduced to why, for so doing would disregard the subordination to 9a. On the other hand it would not do to unite:



because then the subordination of 10 to 9b is ignored. A better notation would be desirable, but the ones that have been conceived to date differ only superficially from the one adopted in the graph.

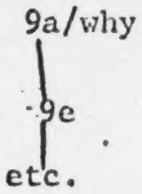
A further point is that, for brevity, the two bottom nodes of the 9a subgraph have been truncated in 9b.

9e. A minor point about implied definitions arises here. The 9eii term is source is nearby star. The text statement was actually a paraphrase of this, namely reflection from nearby star. There seems no need to represent in the graph all paraphrases derived by using common knowledge. As to which one member of the paraphrase set should be represented in the graph, a criterion of adherence to the surface structure would dictate that 9e be replaced by:

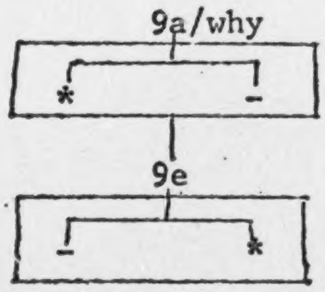

9e' Source, reflection[from source]
source
reflection[from nearby star].

(By the way, the graph carries no implication about the visibility, etc. of the star itself) .

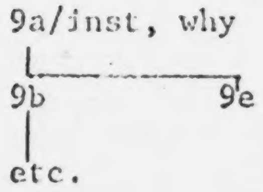
The relation between 9a and 9e deserves comment. In an earlier version of the graph, 9e(=has source) was subordinated to 9a(=has visibility) by



At first thought, this seems plausible, since having a light source is the reason why objects are visible. However, deeper analysis makes this representation appear dubious. As noted in note 9, the predicate has a source is stated explicitly only for the star nebulae. By contrast, the subgraph 9a is stated explicitly only for the cloud nebulae. Hence expanding the above diagram by including subgraphs would give something like (omitting inessential details)

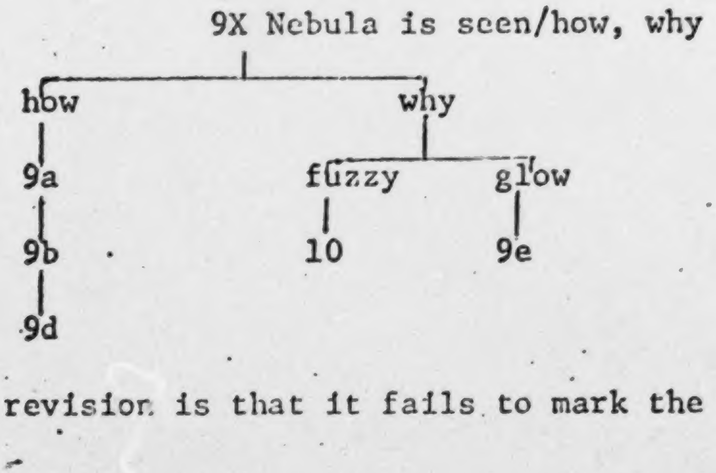


where the * denotes a non-null entry. In order to remove the anomalous alternation of null symbols, one would have to insert other "common knowledge" e.g. that individual gas and dust particles (like individual stars of star nebulae) are not discriminated. The tentative decision to exclude such information rules out this alternative. An even more compelling refutation of the revision is the fact that having a source does not explain why the overall nebula is seen (9ai) but the components are not seen (9aii). If adopted, however, the revision would not incur any other difficulties, since 9b etc. can be attached via:



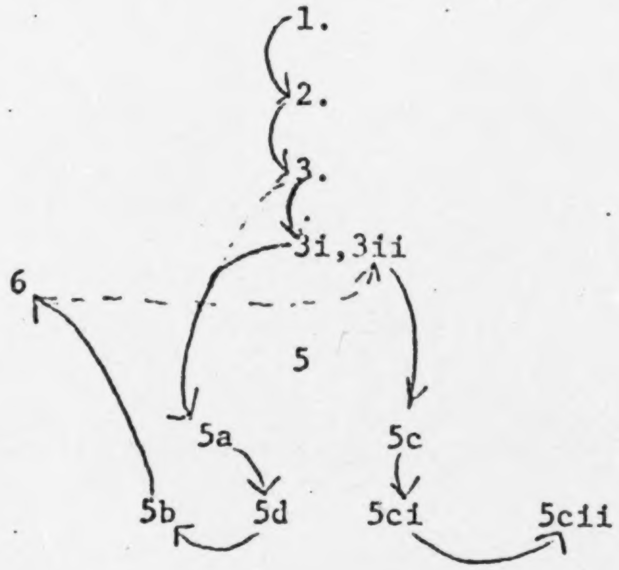
10. It is an interesting question as to whether or not 10(=*Nebula is distant*) should be regarded as inferred from 3(=*Nebula has location relative to our* galaxy). Both the passage and the graph obscure such an inference, yet presumably a speaker would make it. However, the antecedents of 10, namely 9d, 9b, etc. cannot be inferred from 3. Perhaps, therefore, a path directly from 3 to 10 is required. A more radical revision would be to insert 10 between 2 and 3, the subordinator after 10 being *why*. But it is not clear how the 9c subgraph of 10i would get included. This revision probably merits a more detailed examination.

A separate point is that 9 and 10 cannot be rewritten as coordinates by inserting a node 9X thus (omitting subgraphs):

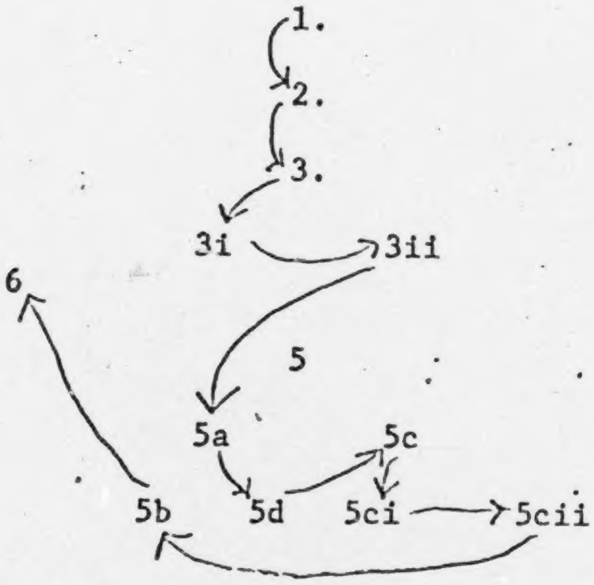


The shortcoming of this revision is that it fails to mark the fact that 10 answers the *why* of 9b.

While a serious discussion of what we are treating as surface properties of the paragraph is outside the scope of the present paper, a brief illustration of the relation between deep structure and sentence order will help clarify the distinction. Diagrams a and b below are excerpts from the nebula structure:



a.



b.

Arrows indicate the sequence of sentences. Thus the diagrams represent two different paths through the same structure, corresponding to the two surface structures. Implied sentences are not on the path, but their subordinates (or recursively, depending on the depth of inference) are on the path. For a given deep structure there will be numerous surface structures, corresponding to varying degrees of explicitness of the text. By the same token, another way of creating different surface structures is the device of repeating sentences for emphasis.

References

Bobrow, D. G. Natural language input for a computer problem-solving system. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass.: M.I.T. Press, 1968. Pp. 135 - 215.

Frase, L. T. Structural analysis of the knowledge that results from thinking about text. J. ed. Psychol. Monogr. 1969, 60 (Part 2), Pp. 1 - 16.

Harris, Z. S. Discourse analysis reprints. Papers on formal linguistics, 1963, 2.

Jacobson, S. N. A modifiable routine for connecting related sentences of English text. In P. L. Garvin and B. Spolsky (Eds.), Computation in linguistics. Bloomington, Ind.: Indiana Univer. Press, 1966. Pp. 284 - 311.

Johnson, R. E. Recall of prose as a function of the structural importance of the linguistic units. J. verb. Learn. verb. Behav., 1970, 9, Pp. 12 - 20.

Koen, F., Becker, A., and Young, R. The psychological reality of the paragraph. J. verb. Learn. verb. Behav., 1969, 8, Pp. 49 - 53.

Quillian, R. The teachable language comprehender: a simulation program and theory of language. Bolt, Beranek, and Newman Technical Report. No. 1963. Cambridge, Mass.: 1969.

Footnotes

[1] This paper was prepared for a Research Workshop on Cognitive Organization and Psychological Processes sponsored by the Committee on Basic Research in Education, National Research Council. The research was supported by Grant No. DBS - 0224 from the U. S. Office of Education and Grant No. MH 15956 from the National Institute of Mental Health.

[2] Thus the simplest grammar for a semantic structure would contain rules for subordination and assignment of values to variables. The rule schema would be:

$A \rightarrow B + C + \ldots$

$A \rightarrow A1 + A2 + \ldots$

and recursively with B or A1, etc. replacing A. Each particular paragraph would be associated with a particular ordered set of such rules. Additional kinds of rules would be required to include the other properties (e.g. correspondences between subgraphs) mentioned in this paper.

[3] In the texts examined to date, the practice of marking correspondences to a standard subgraph (here, the 1 subgraph) determines most of the correspondences between other subgraphs. It is unusual that extra notation is needed to complete the description, as here the 6i symbol describes the correspondence between 6 and 8.

Cognitive Structures and Judgment

by

Roy G D'Andrade

University of California, San Diego

Paper prepared for C.O.B.R.E. Research Workshop on
COGNITIVE ORGANIZATION AND PSYCHOLOGICAL PROCESSES
August 16-21, 1970 Huntington Beach, Calif.

Abstract

This paper attempts to demonstrate that one general procedure used by social scientists to discover how human behavior is organized does not give valid results. This procedure, in which subjects are rated from memory on various attributes, lacks validity because it relies on humans as judges, and human judgment, especially when based on long term memory, is subject to distortion or bias in the direction of pre-existing cognitive structures. As evidence that this procedure is invalid, correlations found between ratings of traits of behavior, where the ratings are based on long term memory, will be shown to be quite similar to judgments about how much alike the terms for these traits are in meaning, but quite different than correlations for these same behaviors which are based on data using the immediate recording of on-going interaction. Implications of these findings for social science theories of human behavior which assume multi-behavior units will be discussed, and an alternate theory proposed.

Humans as Behaviorscopes

In the social sciences people have been used as measuring instruments for the investigation of many aspects of behavior. To find out if the subjects in an experiment are acting aggressively, for example, an investigator will typically use other people as behaviorscopes to observe the subjects. By some unknown and invisible process, these observers will judge whether various segments of the subjects' behavior are aggressive or not. (Specifying what is meant by aggressive by defining the term with reference to more specific acts, such as hitting or insulting, does not really make explicit the decision process, since the rules for coding behavior still depend on a set of undefined terms. No matter how detailed the investigator makes the coding rules, at some point the specifications always fall back on undefined terms, for reasons similar to those discussed below concerning difficulties in measuring behavior mechanically.)

The advantages of methods which use human judgment as a means of measurement appear to outweigh the disadvantages. Of course there are some social scientists who require all their measurements to be solely in units of mass, time, and distance, mechanically recorded. However, if the behavior to be recorded is even slightly complex, it is enormously difficult to try to mechanize the measuring process. The difficulty involves not just problems of detecting small differences in the physical signal, but also very complex problems caused by the lack of one-to-one correspondences between the properties of the physical signal and the distinctions perceived by the human observer. For example, in the area of speech there are at least three quite different acoustic signals which are all heard as initial 'd', depending on the following vowel (Liberman, Cooper, Shankweller, Studdert-Kennedy, 1967).

These many-to-one and one-to-many types of correspondence, which make mechanical decoding of complex human behavior so inefficient and costly, occur between many levels besides the physical media and the perceptual response. For example, there are many ways to indicate agreement besides the sentence 'I agree'. There are alternatives in vocabulary, syntax, and the use of a medium, such as speech rather than gesture or writing. A mechanical device which could determine when a respondent agreed with a statement, whatever the means of agreement used, as well as a human judge can, would be a most amazing and complex piece of machinery. At present, it is much easier to use humans as behaviorscopes than to try to build behaviorscopes from wire and glass.

The Validity and Reliability of Human Behaviorscopes

The low cost of using human judges to assess attributes of human behavior is partially offset by a number of disadvantages. Traditionally, these disadvantages have been discussed as problems of reliability and validity. Reliability is usually thought of as the degree to which measuring instruments of the same type give the same results. A large literature has developed concerning the causes, effects, and remedies for deficiencies in reliability. In general, present-day social science shows a high degree of sensitivity to problems of reliability.

In contrast to problems of reliability, problems of validity are not presently well formulated. In a classic article, Campbell and Fiske presented one of the first methodological articles on ways to assess validity. (Campbell and Fiske, 1959). They argue that validity, like reliability, involves agreement between measures. Validity, however, involves agreement between maximally different, or independent measurement procedures, while reliability involves agreement between maximally similar measures. Thus two different methods which give very divergent results cannot both be valid measures of the same thing, although both might be valid measures of different things. From the sample of examples in the Campbell and Fiske article, it appears that the social sciences have very few methods which show this type of convergent validity in measuring any aspect of human behavior.

Obviously, high reliability alone does not insure validity. It is easy to show that under certain conditions highly reliable judgments can be quite invalid. Optical illusions, like the Muller-Lyer illusion, provide a simple example of conditions which elicit reliable but invalid judgments about which line is longer. The Rosenthal effect, in which subjects reliably report what they were led to hope and expect would happen, rather than what did happen, is a kind of illusion too, but of a cognitive rather than perceptual kind.

This paper attempts to demonstrate that one of the methods used in the social sciences is invalid because it is subject to a special effect analogous to that of an illusion; an effect in which there is a reliable and systematic distortion of certain judgments under special conditions. The general procedure in which this method is often used, and in which the distortion introduced biases the results most drastically, can be described as follows: first, one or more human observers are asked to judge, by ratings or rankings, one or more subjects (possibly including the observers), based on the observer's long term (i.e., over ten minutes) memory of the subjects' behavior, on a number of different traits or categories of behavior. Second, scores are computed for each subject for each trait, usually by taking the mean of all the scores given to each subject on each trait. Third, in order to find out how the traits are related to each, some measure of association, such as the product moment correlation, is computed from the subjects' scores for all pairs of traits. Finally, the measures of association are analyzed to determine how the traits are organized with respect to each other. A number of kinds of analysis may be used, ranging from simple clustering by inspection to factor analysis or multidimensional scaling. The results of such analyses indicate which traits tend to go together, and the overall similarity structure of the trait measurements is taken as a representation of the structure or organization of the subjects' behavior.

The motivation for using this procedure is usually to simplify the description of human behavior by grouping similar traits into multi-behavior units, such as clusters or dimensions. Since human behavior can be described in a very large number of ways, and since there appears to be a wide variety of behaviors which people perform, some procedure is needed for constructing economical descriptions of a person's or a group's behavior. The reason for using ratings or rankings which are based on long term memory in this procedure appears to be that such judgements are very easy to obtain. One can consider the memory of each observer to be like a store house, where the impressions left by hundreds, or even hundreds of thousands of past acts of the subject may be inexpensively and quickly recovered.

Unfortunately, it appears that an observer's memory cannot be considered trustworthy for certain judgments. The evidence to be presented below indicates that there is a systematic distortion in such judgments, in that traits which the observer considers similar will be recalled as applying to the same person, even when this is not the case. As a result of this effect, the correlations found between traits turn out to be due more to the observer's conception of 'what is like what' than to covariation in the behavior of the subjects.

In an earlier paper on this topic it was argued that procedures which try to classify behavior from data consisting of judgments by human observers do not yield information about which behaviors of the subjects go together, but only information about which behavior terms are semantically similar (D'Andrade, 1965). The evidence used to support this contention was a demonstration that in at least two cases judgments of the semantic similarity of trait terms gave approximately the same results as the analysis or correlations based on the ratings of subjects behavior made by informed observers. This evidence turned out to be equivocal, however, since it can be argued that the semantic similarity of trait terms corresponds to the way these traits actually go together, perhaps because the actual relations between traits become semantically coded into the trait terms. Given this isomorphism hypothesis, the fact that the same type of similarity structure can be obtained from both judgments of semantic similarity and from observer ratings would be expectable, and certainly would not invalidate psychological theories which assume that people can be accurately described in terms of multi-behavior units of some kind.

The Multitrait-Multimethod Matrix Correspondence Technique

In order to decide between the isomorphism hypothesis and the systematic distortion hypothesis, a comparison between ratings based on long term memory and the actual behavior of the subjects is needed. If the observer's memory based ratings showed a very different pattern of correlations than the pattern of correlations found for the data based on the actual behavior of the subjects (but a pattern similar to judgments of semantic similarity), then it would be reasonable to not accept the isomorphism hypothesis, and to consider the systematic distortion hypotheses supported. This strategy is similar to the Campbell and Fiske multitrait-multimethod matrix technique, except that instead of comparing specific correlation coefficients, entire patterns of correlation coefficients are to be compared. Pattern comparison is needed here because it is not the validity of specific traits or categories which is in question, but the validity of the correlations found between traits.

Ideally, in order to record the actual behavior of the subjects, a mechanical device should be constructed to count frequencies of different kinds of behavior. Unfortunately, for all the reasons discussed above, a mechanical measuring instrument is not presently practical for any judgment more complex than is talking versus is not talking. The closest approximation to a mechanical device appears to be a trained observer using a simple coding scheme to record a subject's behavior as it occurs. The immediacy of the observer's assessment and the simplicity of the coding decisions should, it is hoped, protect against systematic distortion of the type thought to take place in long term memory.

The most frequently used coding scheme for recording on-going behavior is the Bales Interaction Process category system. The Bales categories are relatively simple in terms of coding rules, and are also useable for ratings or rankings based on long term memory. A schematic classification of the Bales system is presented in Figure 1.

A: Positive Reactions	1. Shows solidarity, raises others' status, jokes,gives help, reward 2. Shown tension release, shows satisfaction, laughs 3. Agrees, shows passive acceptance, understands, concurs, complies
B: Problem Solving Attempts	4. Gives suggestion, direction,implying autonomy for other 5. Gives opinion, evaluation, analysis, expresses feeling, wish 6. Gives orientation, information, repeats, clarifies, confirms
C: Questions	7. Asks for orientation, information, repetition, confirmation 8. Asks for opinion, evaluation, analysis, expression of feeling 9. Asks for suggestion, direction, possible ways of action
D: Negative Reactions	10. Disagrees, shows passive rejection, formality, withhold help 11. Shows tension increase, asks for help, withdraws "Out of Field" 12. Shown antagonism, deflates others' status, defends or asserts self

FIGURE 1

Bales Interaction Process Analysis Categories

(Taken from Parsons and Bales, Family, Socialization and Interaction Process, 1955, p. 267.)

The Borgatta, Cottrell and Mann Study

Because of the general applicability to different methods of measurement of the Bales category system, a search was made to find published studies in which both immediate recording of on-going behavior and judgments based on memory using the Bales categories had been carried out. Two studies were found in which the Bales categories were used for both these methods of measurement, and in which the data were reported completely enough to permit the necessary multitrait-multimethod matrix comparisons. The first study was conducted by Borgatta, Cottrell and Mann (1958). The second study was conducted by Richard Mann, and reported in his Ph.D. thesis (1959).

As part of the Borgatta et al study a number of small groups were observed directly, with immediate recording of on-going behavior according to the Bales category system. After twenty hours of small group contact the individuals in each group also ranked each other on a variety of personality traits, including a set of behavior descriptions which correspond approximately to the Bales category system.

The subjects in the Borgatta et al study consisted of 47 graduate students enrolled in a class on interpersonal relations. The subjects were divided into five small groups, each meeting for a two-hour discussion period every week throughout the semester. The groups had little or no external supervision. The primary focus of discussion was the analysis of processes that go with 'democratic leadership'.

The questionnaire ranking data was collected by the course instructor after the ninth week ostensibly as part of a 'student evaluation procedure'. Rankings were made on a total of 40 traits, with each person ranking all members of his group. Since the groups were of slightly different sizes, rank scores for each group were equalized. The traits which most closely approximate the Bales categories were:

1b. 'shows solidarity and friendliness' (Bales #1)
2b. 'is responsive to laughter' (Bales #2)
4b. 'makes the most suggestions' (Bales #4)
10b. 'disagrees most' (Bales #10)
11b. 'tends to be nervous' (Bales #11)
12b. 'tends to be antagonistic' (Bales #12)

Presumably the investigators felt their re-phrasings of the category labels made it easier for the subjects to rate each other.

The immediate recording of on-going interaction was carried out by a trained observer using the Bales category system. Each of the five groups was observed during the ninth and tenth weekly sesssions. Interaction scores were adjusted to the individuals' rate of initiated acts per 100 minutes. Generally the groups show slightly different and stable interaction profiles.

A factor analysis of the individual rank scores for the forty traits was performed, and two main factors, an 'assertiveness-dominance' dimension and a 'solidarity-friendliness' dimension were found. These results are not directly relevant to the present issue, but do replicate the ubiquitous 'Leary-grid' pattern found when data based on long term memory judgments is used. This Leary-grid organization of interpersonal behaviors, it should be mentioned, can also be obtained using only judgments of semantic similarity (D'Andrade, 1965).

In order to relate trait rankings to the results of the immediate recording of on-going behavior, product moment correlations were computed by Bogatta et al between all variables. These correlation coefficients incicate the degree to which individuals who have a high score on one variable also have a high score on the other variable.

From the complete matrix of correlations it is possible to compare the pattern of correlations found between the Bales categories scores for rates of behavior based on immediate recording to the pattern of correlations found between Bales category scores for ranking based on long term memory. However, since Borgatta et al did not obtain any measures of semantic similarity, in order to make a comparison between the pattern of semantic similarity occurring between the Bales categories and the patterns of correlations found for the other types of data, a test of semantic similarity had to be constructed and administered.

Because the wording of the trait descriptions used by Borgatta et al are slightly different than in the original Bales category system, a modified set of behavior descriptions were constructed for the semantic similarity test which are somewhat closer to the original Bales wording. The phrases used in the semantic similarity test were:

1s. 'shows solidarity' (Bales #1)
2s. 'jokes, laughs' (Bales #2)
4s. 'suggests, gives direction (Bales #4)
10s. 'disagrees' (Bales #10)
11s. 'shows tension, nervous' (Bales #11)
12s. 'shows antagonism' (Bales #12)

The test was administered in questionnaire form to ten graduate student raters, none of whom were previously acquainted with the Bales category system.

The format used for this questionnaire is presented below:

JUDGMENT OF BEHAVIORAL DESCRIPTIONS

The purpose of this questionnaire is to obtain your judgment about the degree of similarity between a series of descriptions of human behavior. Base your judgments of similarity on the meanings of the terms and the degree to which the descriptions typically refer to the same kinds of behavior.

Each question on the test consists of a pair of behavior descriptions to be judged on a seven point scale ranging from 'very similar' to 'very dissimilar.'

In the example below, please rate the descriptive terms 'cooperative' and 'helpful' according to your estimate of their degree of similarity.

cooperative :: helpful

very similar	generally similar	slightly similar	unrelated not similar not dissimilar	slightly dissimilar	generally dissimilar	very dissimilar
+3	+2	+1	0	-1	-2	-3

Place an X in the slot which corresponds most closely to your judgment of how similar the descriptive term 'cooperative' is to the term 'helpful'.

Now go on to the rest of the test. Please answer all the questions.

1. shows solidarity :: suggests, gives direction

very similar	generally similar	slightly similar	unrelated not similar not dissimilar	slightly dissimilar	generally dissimilar	very dissimilar
+3	+2	+1	0	-1	-2	-3

Table 1 below presents the correlation matrices from Borgatta et al for the relevant Bales categories for rates of behavior based on immediate recording, for rankings based on long term memory, and a matrix for the mean semantic similarity ratings (ranging potentially from -3.0 for 'very dissimilar' to +3.0 for 'very similar').

TABLE 1

Comparison of Correlation matrices for Behavior Rates based on Immediate Recordings, Rank Judgments based on Long Term Memory, and Semantic Similarity Ratings, all using Bales Interaction Process Analysis Categories.

Correlations for

Behavior Rates

	1a	2a	4a	10a	11a	12a
solidarity	1a	.35	.35	.06	.09	.17
jokes	2a		.22	.08	-.11	-.13
suggests	4a			.17	.04	-.06
disagrees	10a				.08	.13
tension	11a					.05
antagonism	12a					

Rank Judgments

	1b	2b	4b	10b	11b	12b
solidarity	1b	.32	.24	-.33	-.22	-.50
jokes	2b		.34	-.04	.08	-.04
suggests	4b			.35	.00	.17
disagrees	10b				.35	.75
tension	11b					.28
antagonism	12b					

$r_s = .34$

$r_s = .03$

$r_s = .60$

Semantic Similarity Ratings

	1s	2s	4s	10s	11s	12s
solidarity	1s	.0	.2	-.7	-1.4	-1.5
jokes	2s		.1	-.9	.3	-1.0
suggests	4s			-1.1	-1.4	-1.1
disagrees	10s				.3	.7
tension	11s					1.0
antagonism	12s					

(r_s coefficients are Spearman's rank order correlation measure)

To compare matrices the agreement in rank order for corresponding coefficients was used as a measure of over-all pattern similarity. Degree of agreement in rank order is given by the Spearman rank order correlation coefficients. These coefficients indicate how well the degree of similarity found between a given pair of Bales categories for one method of measurement predicts the degree of similarity to be found between the same pair of categories for a second method of measurement. Thus the coefficients inside a matrix indicate how much alike the Bales categories are for a particular method of measurement, while the rank order coefficients between matrices indicate the degree to which different methods of measurement yield similar patterns of association between Bales categories.

The results presented in Table 1 may be summarized as follows:

(1) Correlation coefficients for behavior rates based on immediate recording tend to be smaller in absolute size than the correlation coefficients for rank judgments based on long term memory.

(2) The correlation matrix for behavior rates is not strongly similar to the correlation matrix for rank judgments (r_s = .34).

(3) The matrix of semantic similarity ratings is fairly similar to the correlation matrix for rank judgments (r_s = .60), but not similar to the correlation matrix for behavior rates (r_s = .03).

Generally, information about which Bales categories are most alike using behavior rates based on immediate recordings will not predict very well which Bales categories are most alike when using rank judgments based on long term memory. The pattern of association for the rank judgments for the Bales categories is more like the pattern of semantic similarity for these categories than it is like the pattern of association found for the observed rates of these behaviors.

These results appear to support the hypothesis that people tend to recall traits considered similar as characteristic of the same individuals. The evidence from the behavior rates based on immediate recording does not support the isomorphism hypothesis, since the categories considered similar in meaning are not the categories which show strongly correlated behavior rates.

What seems to happen in this type of small group situation is that while group members agree to a certain degree with each other in their assessments, a number of these assessments are not based on fact, but on a combination of other assessments. For example, according to the behavior rate correlations individuals who show solidarity are very slightly more than not likely to also show antagonism (r = .17). But since people conceive of 'solidarity' as very dissimilar to 'antagonism' (semantic similarity rating of -1.5), if they notice that an individual shows a good deal of solidarity in his behavior, he will be remembered as not showing antagonism (rank judgment based on long term memory r = -.50).

Further information about the relation between the behavior rates and the rank judgments can be obtained from an inspection of the correlations between these two sets of scores. These correlations, presented in Table 2, are based on the scores assigned to individual subjects; a high correlation coefficient between a particular behavior rate category and a particular rank judgment category means that given an individual's score on one measure a fairly good prediction can be made of that individual's score on the other measure. In Campbell and Fiske's terminology, Table 2 presents a hetromethod correlation matrix, and the correlations between categories with the same label are validity coefficients.

TABLE 2

Correlations Between Behavior Rate Category Scores and Rank Judgment Categories Scores

Behavior Rate Category		Rank Judgment Category					
		1b	2b	4b	10b	11b	12b
solidarity	1a	.23	.48	.36	.31	.03	.22
jokes	2a	.00	.17	.14	.17	.03	.17
suggestion	4a	.08	.17	.38	.09	.06	.00
disagrees	10a	-.03	.05	.11	.30	.11	.22
tension	11a	.07	.00	.29	.16	.29	.00
antagonism	12a	.17	.43	.20	.08	.06	.14

correlations $\geq \pm .24$ significant at $p \geq .05$ for $n = 47$

While the underlined validity coefficients average slightly higher than the hetromethod-hetrotrait coefficients, the size of these coefficients is not impressive, nor can the overall pattern be summarized simply.

The R. Mann Study

The complex pattern of correlations found in Table 2 raises a possibility which might explain the lack of correspondence between the behavior rate correlations and the rank judgment correlations. This is the possibility that the Bales categories mean something different to the subjects, who make the rank judgments, than they do to the observer who does the coding of on-going behavior. A related possibility is that behavior rate scores are overly influenced by the general activity level of the individual, so that, for example, a person who says very little, but what he does say is often a suggestion, comes to have a lower rate score on suggests than an individual who performs a much wider variety of behaviors, but is extremely active. In contrast to behavior rate scores, the ranking judgments may reflect more the proportions of a person's behavior which fall in the various categories.

To control for these possibilities requires a study in which the category scores for the immediate recording of behavior are given in terms of proportions of the individual's total output of behavior, and in which the observer not only does the immediate recording of behavior, but also makes memory based judgments about what proportion of each person's behavior falls into the various categories.

A study with both these procedures was carried out by Richard Mann (Mann, 1959). The Mann study also has an extra feature in that the small group sessions took place under two different experimental conditions, with all subjects taking part in both types of group sessions. Mann's research, described in his Ph.D. thesis at the University of Michigan, is an excellent example of comprehensive and careful data reportage. It includes test and performance scores for every subject on all variables, complete intercorrelation matrices, and examples of the test forms and instructions.

The major purpose of Mann's study was to test the relationship between personality and small group performance. Based on an exhaustive survey of the literature, Mann selected four major personality dimensions: intelligence, adjustment, extroversion, and conservatism. A variety of questionnaire tests were given to obtain measures on these dimensions. A factor analysis of the results yielded six factors, interpreted as measures of adjustment, reasoning ability, social extroversion, lack of self control, conservatism and verbal intelligence.

Personality measure factor scores were computed for each subject, and these then related to the group performance variables. For purposes of this paper, it is the group performance measures which are of interest. These include the use of a modified Bales category system for the immediate recording of on-going interaction. The same categories were also used for memory based ratings, made by both the small group participants and the observer after the small group sessions.

For this study 100 male undergraduates were enlisted as subjects, all of whom had recently pledged fraternities. After taking the questionnaire tests each subject was assigned to two different five-man groups in which none of the participants were well known to each other, and in which no two subjects were together in both groups. One of the groups in which each subject participated worked on a relatively specific task (the 'mined-road' problem) for fifty minutes, with the promise of a substantial reward if that group came up with the best solution. The other group for each subject worked for fifty minutes on a more emotional and diffuse problem concerning 'the way houses should handle pledge training', attempting to formulate a compromise group policy, and then ranking the five fraternities represented by the group member in terms of their closeness to this ideal policy. Experimental conditions were balanced for order, with half the subjects beginning with the 'Task' condition, and half with the 'Social-Emotional' condition.

Group sessions were held in a social science laboratory, and observed through a one-way mirror. After the fifty minute sessions a post-meeting questionnaire was given to the subjects at separate tables, and the observer also filled ratings on each group member.

The interaction during the fifty minute session was scored using a modified version of the Bales category system. Each act was scored in only one category. The modified system is presented below:

	Modified Categories (Bales-Mann)	Bales Categories
(1)	agreement and solidarity	shows solidarity (#1) except for omission of all joking behavior plus shows agreement (#3)
(2)	laughing and joking	shows tension release (#2) plus all joking behavior
(3)	suggesions	gives suggestions (#4)
(4)	opinions and orientations	gives opinions (#5) and gives orientation (#6)

(5)	questions	asks for orientations	(#7)
		asks for opinions	(#8)
		asks for suggestions	(#9)
(6)	disagreement and antagonism	shows disagreement	(#10)
		shows antagonism	(#12)
(7)	tension	shows tension	(#11)

By grouping the Bales categories into fewer classes Mann hoped to make a simpler and more reliable coding scheme, and to achieve a more normal distribution of categories. (Unfortunately, it is often the case that when using the Bales categories in fifty minutes of interaction that half the group members have no acts in one or more of the less frequently performed categories.) The observer had previously worked with Bales, and achieved reliabilities of approximately .90 on the majority of the categories.

The scores for the immediate recording data were transformed into percentages by taking the subject's act frequencies for each of the Bales-Mann categories and dividing it by the subject's total number of acts.

The questionnaire administered after each session to the subjects contained sixteen rating measures. Each of the five subjects rated the other four on these measures, using a ten point rating scale with no ties permitted. The questions for the ratings on the Bales-Mann categories are presented below.

1. (Agreement) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to agree with what others had said?

2. (Laughing and Joking) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to laugh and joke around?

3. (Suggestions) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to give suggestions about what the group should do next or what decisions should be made?

4. (Opinions and Orientations) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to give his opinion or state the facts about things?

5. (Questions) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to ask questions of other people?

6. (Disagreement) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to disagree with what others said?

7. (Tension) Regardless of how much he talked altogether, how would you rate each member of this group (excluding yourself) on how much he tended to be nervous, tense, or ill at ease?

The score for each individual on each measure was computed by simply summing the ratings given by the other four members.

The observer's ratings were based directly on his estimates of each persons act proportions for the six Bale-Mann categories agreement and solidarity (#1), laughing and joking (#2), suggestions (#4), questions (#5), disagreement and antagonism (#6), and tension (#7). No reason is given for the observer not using Bales-Mann category #3 opinions and orientations. The observer also made ratings on a number of other measures, such as leadership and likability.

A comparison of the Bales-Mann immediate recording category definitions and the wording of the questionnaire ratings shows that the category descriptions were simplified for the questionnaire. Where two of the Bales categories were combined, the questionnaire usually used only the higher frequency category as the basis for the questionnaire rating. Thus agrees (Bales #3) has an average percentage rate of about 9.5, while shows solidarity (Bales #1) has a percentage rate of only 2.3. In the questionnaire, Mann uses the phrase 'tended to agree' as the crucial frame of reference for the Bales-Mann category #1 of agreement-solidarity.

In order to compare the correlations computed from the immediate recording percentages and the questionnaire ratings with semantic similarity ratings for the Bales-Mann categories required that some choices be made in the wording of the category definitions for a semantic similarity test. As in the Borgotta et al study, an attempt was made to stay as close as possible to the wording of the immediate recording categories, but at the same time to follow the technique used in the questionnaire of excluding the less frequent Bales category where combinations of the Bales categories were involved. The phrasing of the Bales-Mann categories for the semantic-similarity test are presented below:

Bales-Mann 1	'agrees with others, complies'
Bales-Mann 2	'laughs and jokes'
Bales-Mann 3	'gives suggestions'
Bales-Mann 4	'gives opinions and states facts'
Bales-Mann 5	'asks questions'
Bales-Mann 6	'disagrees with others, indicates contrary opinion'
Bales-Mann 7	'shows tension and nervousness, withdraws'

Originally a questionnaire similar to the questionnaire used to measure semantic similarity for the Borgotta study was constructed. However, perhaps because of the increase in the number of categories and the somewhat smaller but more complex range of meaning contained in the Bales-Mann categories, the rating test proved relatively unreliable, the results showing a systematic but weak correspondence to the subjects' and observer's questionnaire rating correlations.

A more reliable and sensitive test method was constructed by having all the pairs of Bales-Mann categories typed out cards, and asking respondents to rank order all the pairs of categories.

Instructions were given as follows:

> On each of these cards are typed two somewhat different phrases which describe how people act. Please place the cards in rank order one above the other according to how likely it is that the same behavior might be described by both of these two phrases. The top card should be the card with the two phrases which are most likely to describe the same behavior. The last card should be the card with the two phrases which are most unlikely to be descriptions of the same behavior.

Respondents were tested individually. The more specific and stringent criterion of referential overlap, i.e., the probability that two phrases may refer to the same event, was used instead of the looser criterion of 'similarity in meaning', with the hope of improving reliability.

A total of five respondents, all graduate students or wives of graduate students (none with a social science background) were tested. The results indicate a relatively high degree of consensus concerning the referential similarity of the Bales-Mann categories, with a Kendall's W of .70, which corresponds to an average rank order correlation between subjects of + . 62.

The correlation matrices for immediately recorded behavior, post session subject and observer ratings, for each of the two experimental group conditions, taken directly from Mann, 1959, Tables 2-J and 3-J, Appendix H are presented below in Table 3.

TABLE 3

Correlations for Immediate Recordings, Subject Ratings, and
Observer Ratings using Bales-Mann categories under
Two Experimental Group Conditions

TASK CONDITIONS

Correlations for Behavior percentages from Immediate Recordings

		2	3	4	5	6	7
agree	1	.14	-.26	-.25	.17	-.32	-.13
joke	2	*	-.27	-.34	-.20	.06	.00
suggest	3		*	-.18	-.23	-.11	-.28
orient	4			*	-.27	-.16	-.22
question	5				*	-.20	.24
disagree	6					*	.06
tension	7						*

Correlations for Subject Ratings

		2	3	4	5	6	7
agree	1	.16	-.35	-.41	-.18	-.61	.32
joke	2	*	.31	.34	.28	.16	-.19
suggest	3		*	.83	.53	.49	-.35
orient	4			*	.47	.50	-.46
question	5				*	.45	-.01
disagree	6					*	-.03
tension	7						*

Correlations for Observer Ratings

		2	3	(4)	5	6	7
agree	1	.28	.05	*	.40	-.34	-.32
joke	2	*	.34	*	.16	.05	-.15
suggest	3		*	*	.39	.54	-.25
(orient)	4			*	*	*	*
question	5				*	.09	-.26
disagree	6					*	.12
tension	7						*

SOCIAL-EMOTIONAL CONDITIONS

Correlations for Behavior percentages from Immediate Recordings

		2	3	4	5	6	7
agree	1	.01	.02	-.44	.01	-.30	-.11
joke	2	*	-.37	-.30	-.25	-.18	.02
suggest	3		*	-.33	.37	.25	-.07
orient	4			*	-.33	-.33	-.32
question	5				*	.22	-.06
disagree	6					*	.14
tension	7						*

Correlations for Subject Ratings

		2	3	4	5	6	7
agree	1	.20	.18	-.15	.14	-.55	.04
joke	2	*	.32	.37	.23	.23	-.16
suggest	3		*	.58	.61	.25	-.24
orient	4			*	.39	.43	-.34
question	5				*	.32	-.11
disagree	6					*	.12
tension	7						*

Correlations for Observer Ratings

		2	3	(4)	5	6	7
agree	1	.40	.33	*	.41	-.42	-.48
joke	2	*	.40	*	.41	-.17	-.33
suggest	3		*	*	.63	.35	-.24
(orient)	4			*	*	*	*
question	5				*	.07	-.28
disagree	6					*	.30
tension	7						*

Table 4 presents the mean semantic similarity rankings. (Actually, since the pair judged most similar by each ranker was given a score of 21, and the pair judged least similar was given a score of 1, the figures in Table 4 are technically mean dissimilarity scores. The scores were reversed so that the semantic similarity scores would have the same direction-size relation that correlation coefficients have; i.e., the larger the coefficient the more similar the variables.)

TABLE 4

Mean Semantic Similarity Rankings*

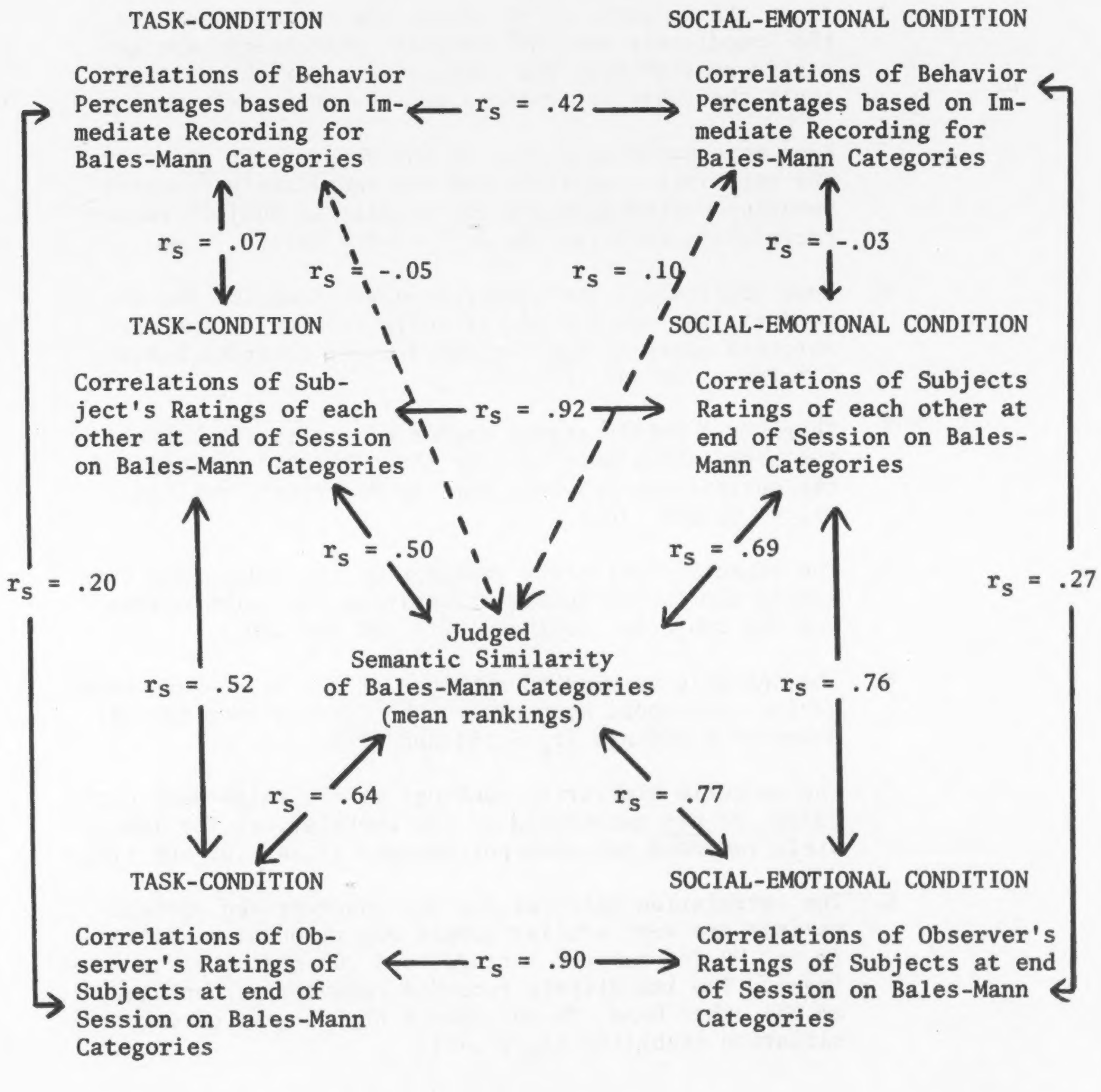
		2	3	4	5	6	7
agree	1	17.0	16.2	11.6	13.6	2.8	4.8
joke	2		12.0	10.2	12.0	5.6	4.8
suggest	3			15.6	19.0	9.6	5.4
orient	4				15.0	10.8	3.2
question	5					14.6	5.2
disagree	6						3.8
tension	7						

* the greater the mean figure the higher the similarity ranking

Table 5 presents the Spearman rank order correlation coefficients for the comparison of the different matrices. As in the Borgatta et al study, a high rank order coefficient indicates that the over-all pattern of correlations in the two matrices are similar, with the Bales-Mann categories which were found to be alike under one condition or method of measurement also found to be alike under the other condition or method of measurement.

TABLE 5

Comparison of Correlation Matrices for Behavior Percentages Based on Immediate Recording, Post-Session Subject Ratings, Post-Session Observer Ratings, each under two Group-Experimental Conditions, with Semantic Similarity Rankings, all using the Bales-Mann Category System


TASK-CONDITION
Correlations of Behavior Percentages based on Immediate Recording for Bales-Mann Categories
r_s = .42
SOCIAL-EMOTIONAL CONDITION
Correlations of Behavior Percentages based on Immediate Recording for Bales-Mann Categories
r_s = .07
r_s = -.05
r_s = .10
r_s = -.03
TASK-CONDITION
Correlations of Subject's Ratings of each other at end of Session on Bales-Mann Categories
r_s = .92
SOCIAL-EMOTIONAL CONDITION
Correlations of Subjects Ratings of each other at end of Session on Bales-Mann Categories
r_s = .20
r_s = .50
r_s = .69
r_s = .27
Judged Semantic Similarity of Bales-Mann Categories (mean rankings)
r_s = .52
r_s = .76
r_s = .64
r_s = .77
TASK-CONDITION
Correlations of Observer's Ratings of Subjects at end of Session on Bales-Mann Categories
r_s = .90
SOCIAL-EMOTIONAL CONDITION
Correlations of Observer's Ratings of Subjects at end of Session on Bales-Mann Categories

The results presented in these tables reinforce the findings reported from the Borgatta et al study, and give additional information concerning the way in which changes in situation affect the different methods. The results may be summarized as follows:

1. As in the Borgatta et al study, the correlations for the immediately recorded behavior percentages are generally smaller than the correlations for the ratings (both the subjects' ratings and the observer's ratings).

2. Even more strikingly than in the Borgatta et al study, the correlation matrices for the immediately recorded behavior percentages are not similar to subject rating correlation matrices (r_s = .07 and -.03).

3. Most critically, the correlation matrices for the observer's ratings are not strongly similar to the correlation matrices for the immediately recorded behavior (r_s = .20 and .27).

4. There is a fairly strong degree of similarity between the correlation matrices for the subjects' ratings and the correlation matrices for the observer's ratings (r_s = .52 and .76).

5. The semantic similarity rankings of the Bales-Mann categories correspond fairly strongly to the correlations for the subjects' ratings (r_s = .50 and .69).

6. The semantic similarity rankings of the Bales-Mann categories correspond strongly to the correlations for the observer's ratings (r_s = .64 and .74).

7. The semantic similarity rankings of the Bales-Mann categories do not correspond to the correlations for immediately recorded behavior percentages (r_s = -.05 and .10).

8. The correlation matrices for the observer and subject ratings are very similar across experimental conditions (r_s = .92 for subject ratings, and .90 for observer ratings). The immediately recorded behavior percentages, on the other hand, do not show a high degree of cross-situation stability (r_s = .42).

The over-all pattern of results shows a consistent network of congruence between the semantic similarity rankings and the correlation matrices for ratings made on the basis of long term memory, regardless of whether the ratings were made by subjects or the observer, or under group task conditions or group Social-Emotional conditions. In contrast, the correlation matrices for the immediately recorded behavior percentages are not strongly similar to any of the kinds of ratings made on the basis of long term memory, and show considerable change across group conditions.

The hetromethod correlation matrices for immediately recorded behavior percentages, subject ratings, and observer ratings, taken from Appendix G of Mann 1959 are presented in Table 6.

TABLE 6

Cross-method correlations for Bales-Mann Categories under two experimental group conditions

TASK CONDITION

Immediately Recorded Behavior Percentages

Subject Ratings

		1	2	3	4	5	6	7
agrees	1	.41	.29	-.44	-.21	.31	-.30	.30
jokes	2	.07	.37	-.17	.16	-.10	.10	-.13
suggests	3	.01	-.23	.20	.42	-.14	.03	-.39
orients	4	.06	-.17	.20	.54	-.16	-.01	-.37
questions	5	.04	-.20	.03	.10	.10	.03	-.21
disagrees	6	-.30	-.26	.07	.29	-.06	.32	-.15
tension	7	.19	.16	-.22	-.07	.59	.13	.55

Observer Ratings									Subject Ratings						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7
agrees	1	.34	-.01	-.05	(.08)	.14	-.49	-.29	.26	.18	.19	(.17)	.30	-.16	-.09
jokes	2	-.00	.24	-.06	(.09)	.09	-.10	-.29	.11	.52	.25	(.32)	.12	.02	-.32
suggests	3	-.19	-.30	.29	(.30)	-.14	.02	-.37	-.40	.28	.64	(.68)	.28	.39	-.42
questions	5	.14	-.32	.05	(.19)	.18	-.29	-.31	-.12	.26	.46	(.42)	.34	.16	-.16
disagrees	6	-.13	-.19	-.02	(.24)	-.10	.51	-.06	-.37	.10	.41	(.38)	.10	.41	-.16
tension	7	-.20	.00	-.38	(.00)	.06	.22	.59	.08	-.23	-.22	(-.27)	.13	-.01	.25

SOCIAL-EMOTIONAL CONDITION

Immediately Recorded Behavior Percentages

Subject Ratings

		1	2	3	4	5	6	7
agrees	1	.36	.27	-.14	.10	-.08	-.53	-.24
jokes	2	.00	.25	-.10	.09	.04	-.00	-.44
suggests	3	.19	-.23	.20	.12	.30	-.16	-.45
orients	4	-.02	-.26	.05	.23	.07	.06	-.49
questions	5	.11	-.18	.04	.08	.36	-.09	-.28
disagrees	6	-.31	-.31	.04	.10	.15	.53	-.07
tension	7	-.16	.15	-.13	.06	-.10	-.02	.14

Observer Ratings									Subject Ratings						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7
agrees	1	.30	.09	-.05	.08	.01	-.41	-.46	.24	.16	.26	(.28)	.21	-.16	-.22
jokes	2	.25	.33	-.06	.08	.12	-.10	-.30	.18	.33	.28	(.24)	.15	-.09	-.25
suggests	3	.12	-.21	.33	.18	.30	.12	-.47	-.13	.34	.61	(.48)	.33	.23	-.35
questions	5	.20	-.02	.18	-.07	.48	-.01	-.36	.02	.15	.48	(.32)	.39	.05	-.34
disagrees	6	-.18	-.27	.20	-.20	.18	.65	.07	-.47	.04	.06	(.15)	.01	.51	-.02
tension	7	-.21	.04	.12	-.17	-.05	.35	.55	-.29	-.13	-.22	(-.19)	-.21	.17	.20

The correlations in Table 6 show a slightly more consistent pattern than the hetromethod matrix from the Borgatta et al study (Table 2). Generally the categories for the immediately recorded behavior percentages have their highest correlations with the properly corresponding subject and observer rating categories (underlined coefficients). These validity coefficients for the immediately recorded behavior percentages and the observer ratings are slightly higher than the validity coefficients for the immediately recorded behavior percentages with the subject ratings.

The overall impression from Table 6 is that the memory based judgments of the ratings are affected by what the person did according to the immediately recorded behavior percentage data, and generally affected most in the correctly corresponding categories, but not strongly. The median 'validity' correlation coefficient is between .33 and .36, which indicates that only about 10% of the variance in the rating scores can be accounted for by the immediately recorded behavior percentages. This leaves great room for slippage, or memory drift. And even a relatively small amount of slippage is sufficient to permit a considerable reordering of which categories are most alike, and hence a drastic shift in the results of any higher-order analysis of how these traits are organized into dimensions or clusters.

General Formulation

The argument presented in this paper is not just that there is memory drift when people make ratings or rankings of other people's behavior, but that this 'drift' is systematic, nonrandom, moving in the direction of the rater's conception of 'what is like what'. More abstractly stated, it is hypothesized here that given a series of attributes (such as behavior traits) which can apply to a class of objects (such as other people), that there will be a systematic shift in the individual's recall of which attributes are possessed by which objects, such that the more similar the individuals conception of any pair of attributes, the more likely it will be that the individual recalls both attributes as belonging to the same objects. As a result of this type of memory shift, any attempt to discover how human behavior is organized into multi-behavior units, such as dimensions or clusters, which is based on data consisting of judgments based on long term memory will result in conclusions which reflect the cognitive structure of the subjects.

It might be argued that since the factor analyses and other types of higher order analysis which have been used on these kinds of data have also included the immediately recorded behavior observation data that this biasing effect would be controlled or contained. But, since the correlations from the immediately recorded behavior data tend to be smaller and less stable than the correlations from the memory based ratings or rankings, and since the immediately recorded behavior categories have their highest outside correlations with the correctly corresponding rating categories, the immediately recorded data becomes absorbed by ratings on other kinds of long term memory judgment data, by being 'fit' into a structure determined by the memory based variables.

Another response to the argument presented here is to say that a major problem in the study of behavior involves the way behavior is perceived to be organized, and how persons and events are placed relative to each other within the individual's cognitive structure. The methods of using ratings and rankings based on long term memory then become perfectly appropriate as ways of getting at these cognitive structures. This use of memory based data seems reasonable and realistic. It does not, however, solve questions about how behavior is actually organized.

Rediscovery

After completing a first draft of this paper, I discovered a reference in an older textbook on psychometric methods (Guilford, 1936) to a study done in 1931 by Theodore Newcomb which clearly states the major hypotheses presented here, supported by differences in correlations found between observer ratings based on long term memory and the correlations found between immediately recorded behavior frequencies (Newcomb 1931). The subjects were 30 problem boys sent to a summer camp for five weeks. The boys were under the constant observation of a psychiatrist and six or more trained counselors. A daily record was kept of specific incidents involving 26 categories of behavior for each boy by his own counselor. Some 8,500 incidents involving these categories were also recorded by the experimenter. At the close of the camp period ratings were obtained from each of the seven observers on the frequency of these 26 categories of behavior for every boy.

A mean correlation of .41 was found for the relation between daily record scores for particular behaviors and the ratings for these behaviors. On bases of this evidence it appears that there was a fair but not especially good degree of validity for the specific behavior category ratings. However, the correlations between behavior categories showed a considerable degree of distortion when the results from the two different methods were compared. Intercorrelation of the categories which were thought to make up nine general traits (a total of 112 correlations) yielded a mean correlation figure for the ratings of .49, but only .14 when computed from the daily records. Concerning these results Newcomb states:

> The conclusion may therefore be drawn that the halo effect, inevitable in the ratings, worked in such a way as to cause the rater to rate similarly logically related behaviors... The close relation between the intra-trait behaviors which is evident in the ratings may, therefore, be presumed to spring from logical presuppositions in the minds of the raters, rather than from actual behaviors. (Newcomb, 1931, p. 288) [Italics added]

Similarity

A critical problem in the formulations presented above concerns the lack of specificity in some of the terms, and especially the vagueness of the term 'similarity' when used with reference to an individual's conceptions. In this paper measures of conceptual similarity have been phrased entirely in terms of semantic similarity. Thus the instructions for the questionnaire and the ranking task stressed that the similarity judgments should be based on the degree to which the events named by the terms are actually alike, rather than the degree to which the terms elicit the same associations, or sound the same, etc.

In the earlier paper on trait psychology and componential analysis the theoretical discussion emphasized denotative similarity; that is, the degree to which different terms shared the same distinctive features, or criterial attributes.

Richard A. Shweder, in a study which demonstrates that results similar to those obtained by Bales work with small groups (Bales, 1969) can also be obtained by having respondents sort the relevant variables on the basis of 'similarity in meaning', has pointed out that any assumption that the respondents are actually using overlap in distinctive features as the primary basis for making similarity judgments is unwarranted (Shweder, 1969). Shweder argues that the basis on which the respondents make similarity judgments of this type is the degree to which attributes contiguously go together in making up a symbolic '*behavioral type*', which he considers a special type of learned cultural construct.

Work by other investigators also indicates that a number of different kinds of relationships may be involved in judgments of similarity. Flavell and Flavell (1959) and Flavell and Stedman (1961) have presented evidence that judgments about similarity in meaning are affected by the '*logico-grammatical relationships*' occurring between terms. In the Flavell and Stedman study children and adults judged which of two pairs of terms were more similar in meaning (e.g., *big* - *large* versus *throw* - *ball*). Across a large number of such judgments, a relatively stable and consistent rank order in similarity was found for the various categories of logico-grammatical relationships by approximately age ten. Highest in the rank order of similarity were *synonymous* pairs (big - large, steal - rob), then *similar-dimension* pairs (small - tiny, smile - laugh), then *superset-subset* pairs (bird-sparrow, tree - oak), follow closely by *whole-part* pairs, (bird - wing, shoe - heel), *object-attribute* pairs (lemon - sour, mouse - small) and *common-action-of* pairs (dog - bark, lion - roar).

Further down the scale were coordinate pairs (cow - horse, pipe - cigar), common-action-upon pairs (sweep - floor, chew - gum), common use pairs (dog - bone, farmer - tractor), and part-part pairs (pedal - handlebars, wall - floor). Last was contrast-on-a-dimension pairs (hard - easy, strong - weak). (A control of unrelated words was not used.)

Results such as the Flavell and Stedman rank ordering could not be due to judgments based solely on overlap in distinctive features. Other criteria, closer to sentence substitutability, appear to be involved. In any case, it is somewhat ironic that the problem of not knowing what is happening when people are used as measuring instruments, raised at the beginning of this paper, and given as the reason for certain procedures in the social sciences producing invalid results, returns to plague a later formulation about exactly how these invalid results come about.

Wider Implications and Alternative Schemas

If the argument presented here concerning the lack of validity of procedures for discovering how behavior is organized when these procedures are based on correlations of long term memory judgments is correct then a large number of studies in the social sciences are brought into question. Studies which are based on correlations from memory based check-lists, rating, or interviews are obviously placed in the 'dubious' category. Studies in which the correlations are based on questionnaire responses, in which the respondent answers on the basis of his recollections either questions about himself or his world would also seem to be placed in doubt, although as yet there is no demonstration that the correlations between items on a personality inventory test, such as the California Psychological Inventory, or a child-rearing questionnaire, such as the Parental Attitude Research Instrument can be reproduced by having a small number of raters judge these items with respect to similarity in meaning.

Most of these correlational studies have attempted to show that certain general traits or dimensions of behavior make possible an economic description of personality, or child rearing, or interpersonal behavior. But if the correlations on which these studies rest are primarily an artifact of the raters or the questionnaire taker's cognitive structure, and not a reflection of the real world, then there is little or no evidence that human behavior can be described by large multi-behavior units. What remains is a world in which human behavior is to be described in terms of specific behaviors occurring in specific situations, as Mischell and others have forcefully argued (Mischell, 1968).

A world made up of numerous ungroupable behaviors might seem disadvantageous for any attempt to describe human behavior. Without multi-behavior unit constructs, such as extroversion, assertiveness, intelligence, etc., it might seem as the goal of accounting for a large portion of an individual's behavior with a relatively small number of descriptive terms is not feasible. However, there is an alternative to the multi-behavior unit schema. This alternative is based on the analysis of the distribution of frequencies with which specific behaviors are performed.

Frequency distributions for category systems applicable to human behavior display a very common form or shape, and it is the shape of these distributions which makes possible descriptions which are brief yet account for large portions of behavior. The shape of a frequency distribution for the nominal or unordered classes typical of most behavior category systems is usually drawn as a rank-frequency graph. On a rank-frequency graph the categories are arranged in rank order of frequency along the horizontal axis, with the most frequent category next to the origin, and the actual frequencies (or proportions of the sum total of all categories) scaled along the vertical axis. Obviously, given any set of frequencies, a line through the frequency plots for each rank will decrease monotonically (i.e., each point will be lower than the point to its left). Figure 2 presents an example of a rank-frequency graph, with four possible kinds of rank-frequency relations plotted by the lines lettered 'A', 'B', 'C' and 'D'.

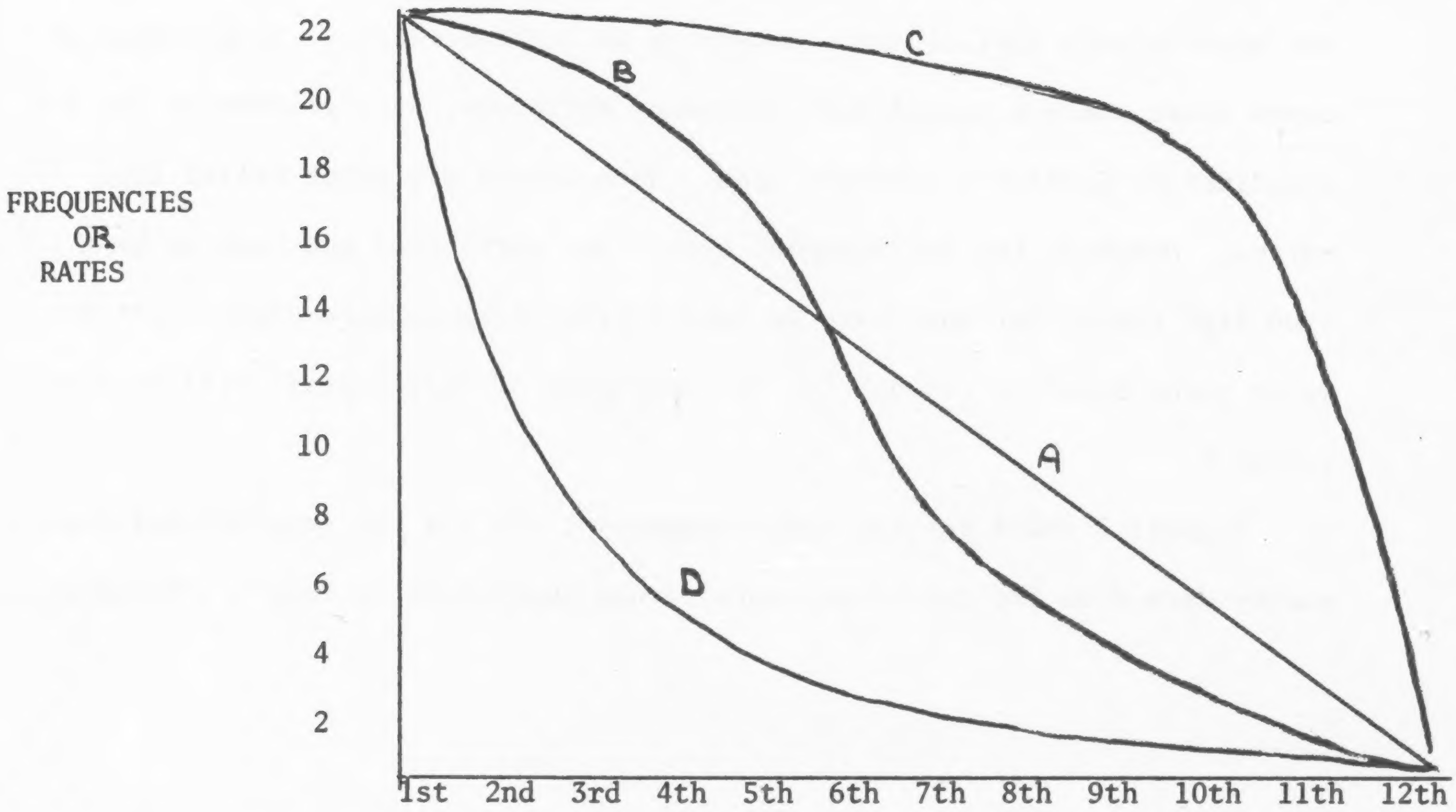


FIGURE 2

RANK-FREQUENCY GRAPH

One of the most general findings in the social sciences is that rank-frequency plots for a variety of classifications of behavior have a shape like curve 'D', with a rapid drop from the first to second ranked category, and successively smaller drops occurring as rank decreases. In a number of cases these concave shaped rank-frequency relations, when graphed on log-log coordinates, display a straight line. Such curves are often called Zipf curves. However, for the argument here, the curve does not have to be a true Zipf curve, but does have to have a pronounced concave shape. (At a later point possible reasons for the concavity of these curves will be discussed.)

Figures 3 and 4 present rank-frequency plots for the proportional-frequency data from the two experimental group conditions of Mann's 1959 study.

FIGURE 3

Bales-Mann Categories for Task Oriented Group

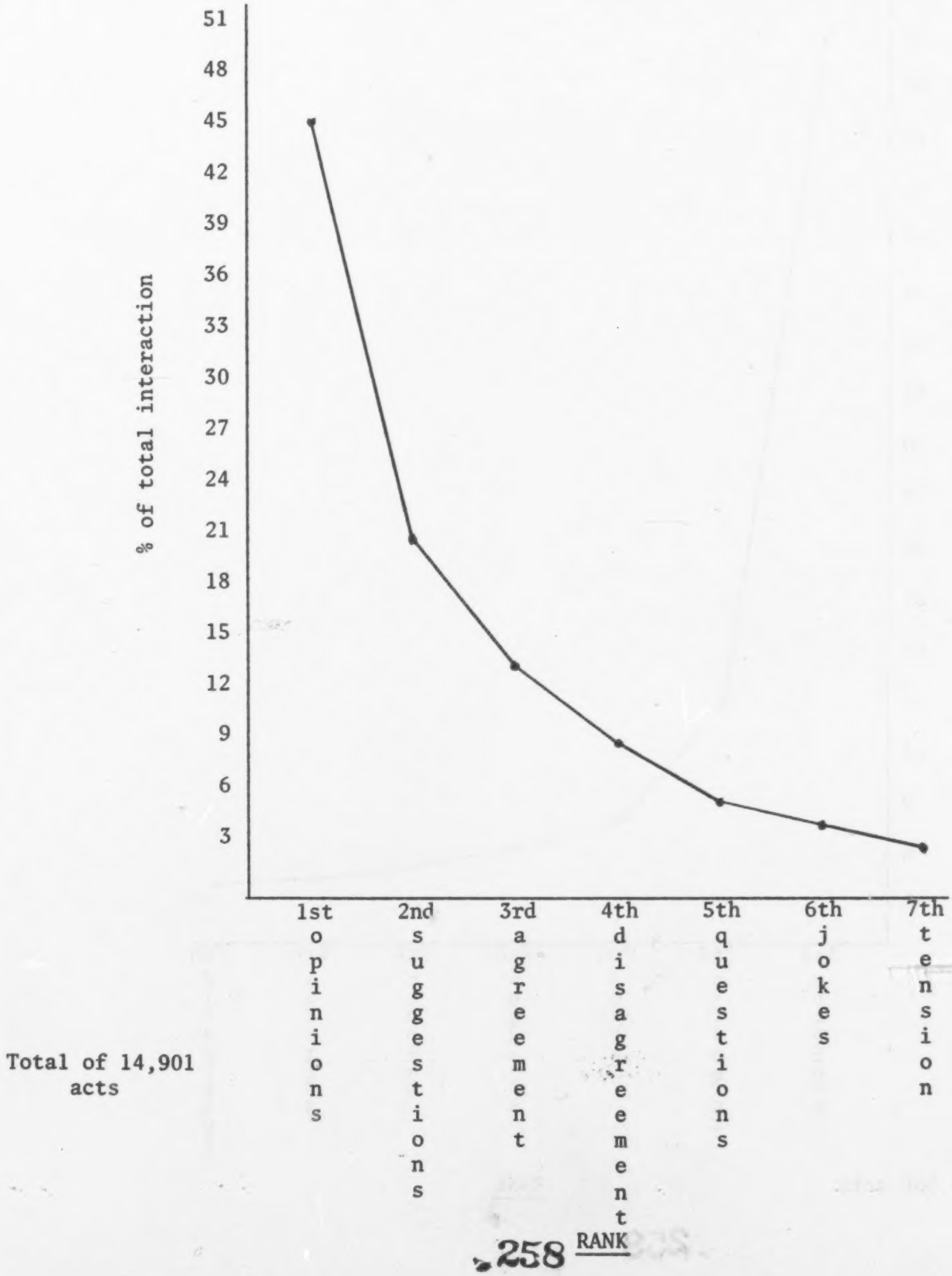
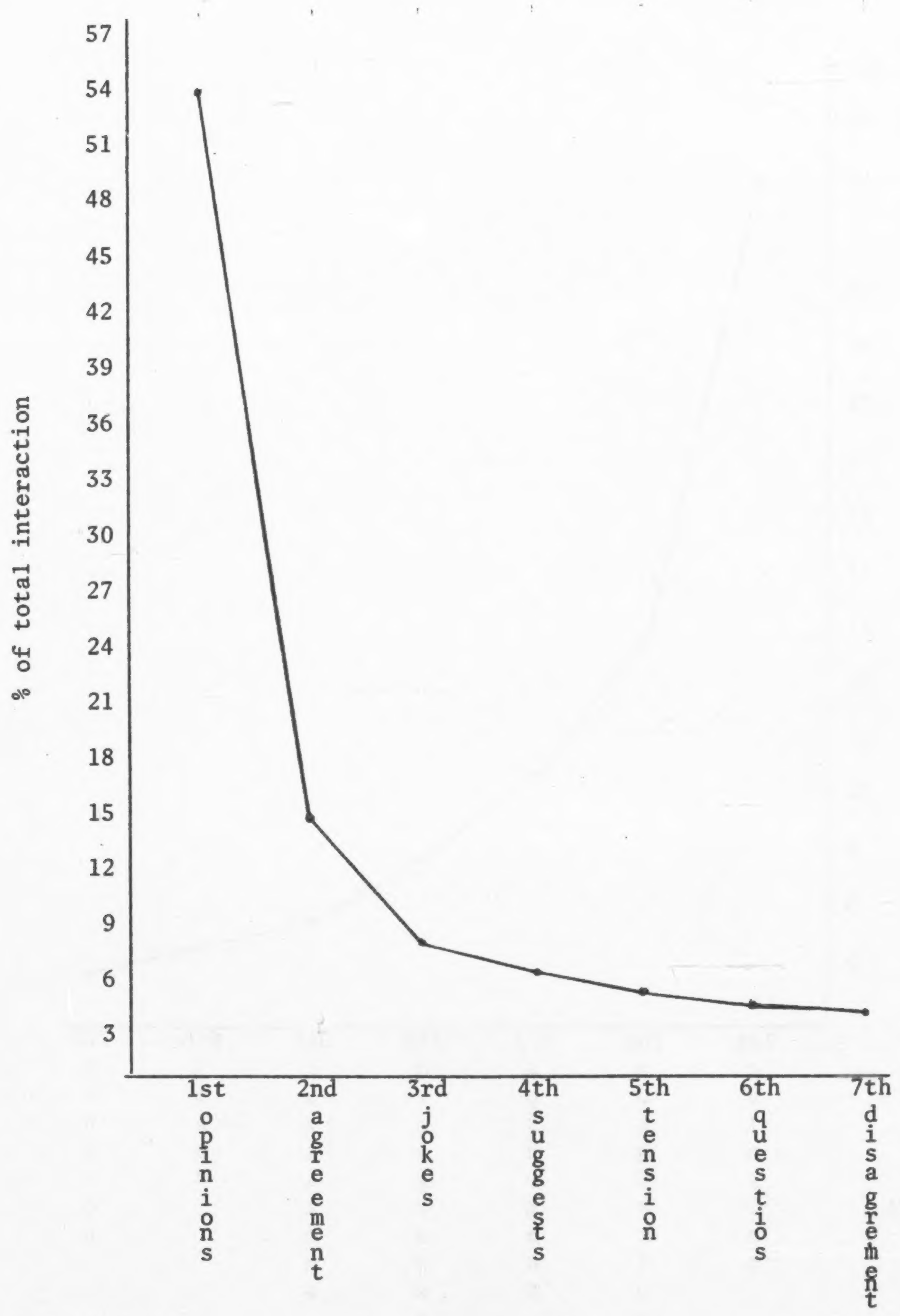


FIGURE 4

Bales-Mann Categories for Social-Emotional Group



Total of 16,550 acts

RANK

Both rank-frequency plots show pronounced concave forms, with the Social-Emotional condition displaying an even steeper initial drop than the task condition. There are other interesting differences between the two groups, with suggests and disapproves moving down in rank order for the Social-Emotional condition compared to the Task condition, while jokes moves up, from sixth to third rank position.

The steep initial drops from the high ranking categories shows that most of the behavior of these small groups occurs in just a few categories. In the task condition group the two categories of opinions and suggestions account for 65% of all the behavior. In the Social-Emotional condition group the two categories of opinions and argument account for over 69% of all the behavior.

In general, concave shape for a rank-frequency curve indicates that a good prediction of what the group or individual is doing can be gained simply by guessing that the group or individual is performing a behavior belonging to a high rank category. Or, to put it another way, the concave curve indicates that the group or individual performs mainly its most frequent behaviors. It should be stressed that if the rank-frequency plot does not have a concave form, then one cannot obtain a good prediction by assuming that an individual is performing his most frequent behaviors. If the shape of the curve is convex, for example, like curve 'C' in Figure 2, then there is very little improvement in prediction to say that the individual is performing his next to last most frequent behavior. (What a convex curve like 'C' indicates is that the individual performs his most infrequent behaviors very infrequently. This kind of situation does not appear to be typical of most people for most behavior category systems.)

The implication of the concave form of rank-frequency relations is that it is not necessary to group behaviors into clusters, or traits, or dimensions, to be able to give an economical description of an individual's behavior. Instead, by taking a number of different category systems, and plotting the rank-frequency relations for each one, a good predictive description of that person's behavior can be given by 'skimming off' the top ranking categories from each classification system, and then describing the person as highly likely to be performing these few behaviors. Thus, if one categorizes a person's gestural repertoire, his paralinguistic repertoire, his durations of speech and silence, his choice of conversational topics, his selection of behavior settings, etc., and simply picks the highest frequency category for each of these classifications, a remarkably predictive description of the person can be obtained.

The critical difference between this behavior specific approach to description and the multi-behavior unit approach is that the behavior specific approach does not assume that because an individual performs one specific aggressive act with high frequency, such as insulting others, that he will perform other aggressive acts, like hitting, or using threat stares, with high frequency. No generality in behavior is assumed, as in the multi-behavior unit approach.

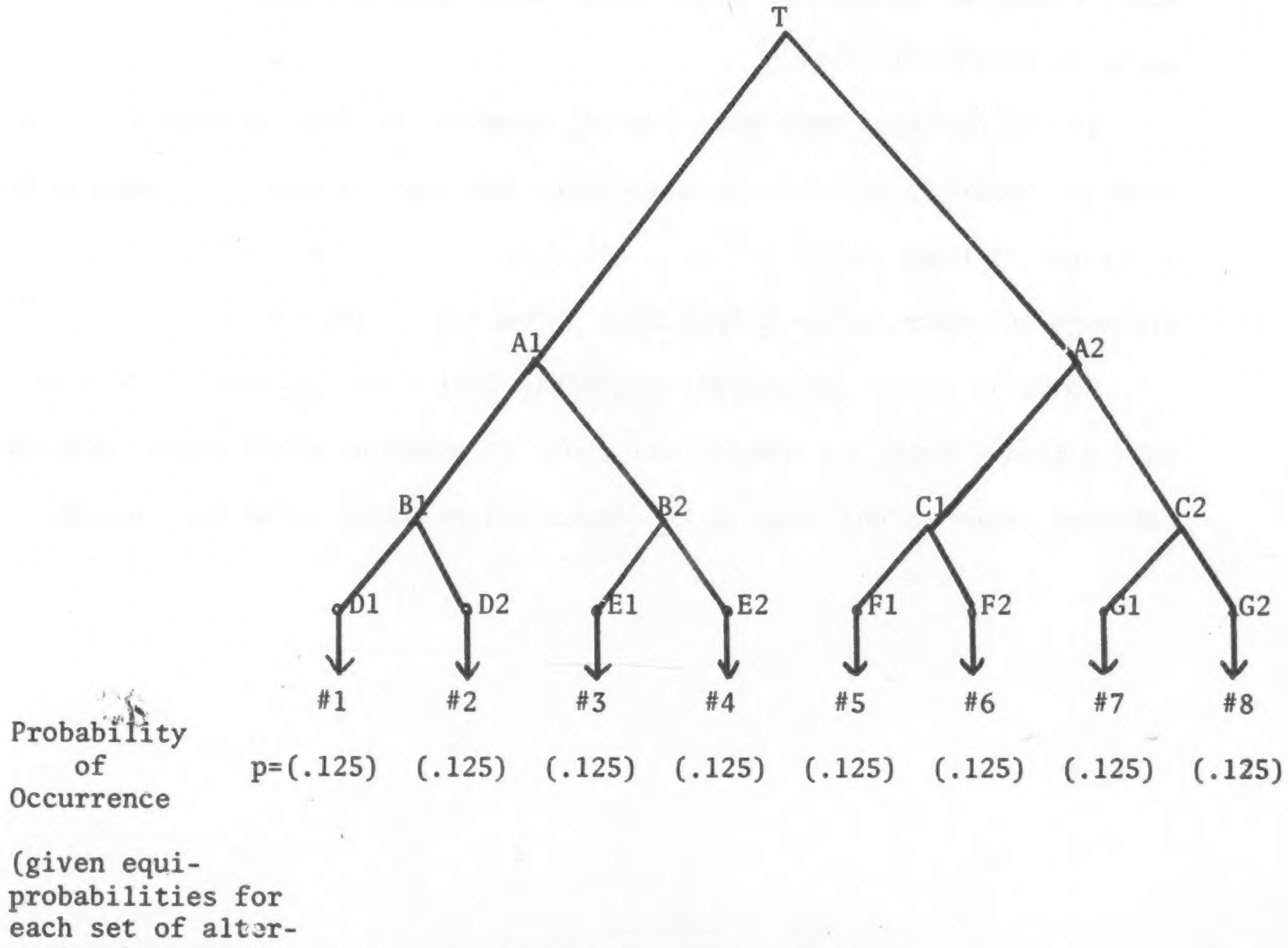
The failure of the multi-behavior unit approach, which assumes generality in behavior, to yield good predictions of individual or group behavior, is, of course, still a moot point in some areas of the social sciences. However, the over-all pattern of evidence seems clear: the conception of humans in terms of dimensions, types, general traits, or any other multi-behavior units has not been found to have reasonable validity (Hunt, 1965, Vernon 1965, Mischell, 1968). With respect to culture and personality studies this means that the description of specific high frequency behaviors said to be typical of a culture are probably quite useful in accounting for a considerable number of behavior occurrences in those groups, while the more abstract descriptions of cultures as Appollonian, or gentle probably lack validity, in that these cultures most probably provide many specific examples of high frequency non-Appollonian or non-gentle behaviors.

Why Concave

The usefulness of describing individuals or groups in terms of their high frequency behaviors depends on the degree of concavity of the rank-frequency relation. The question then arises as to why the rank-frequency relation should almost always show a concave shape for behavior category systems. There would seem to be a number of possible causes. Zipf curves can be neatly generated by a random walk process (Miller and Newman, 1958). It might be speculated that a near random walk process will also yield Zipf-like curves. Since humans have very limited immediate or short term memories, there is a tight upper limit on the possible number of categories in a classification system which involves immediate decisions. If the categories of the classification involved completely cross-cutting decisions, as in the complete parodigm of a kinship system, then the upper limit may exceed 64 categories (Wallace, 1961), but to the extent that the decisions do not cross cut, then the number of categories in the classification reduces to approximately seven or less (Miller, 1956).

Because of these stringent limitations, a human cannot hold in his mind at any one instant the entire collection of behaviors he might perform (e.g., shave, write the letter 'x', yawn, say 'abbracadabra',...) and choose the most appropriate one. Instead, some *sequential* process of decision making must be involved, where the person makes taxonomically high level decisions, selecting from a small set of upper level possibilities (e.g., *get ready*, *get in touch*, *relax*, *perform*), then once the high level choice is made, (or the person is placed in a setting which restricts the choice automatically) the person moves down the taxonomy of choices to the next level of specificity (e.g., *get dressed*, *write Susan*, *stretch*, *do a magic trick*), continuing down the decision tree selecting alternatives until choices about the most specific alternatives which make up the operations of shaving, writing the letter X, etc, are reached.

This sequential process would not necessarily generate Zipf-like curves unless the taxonomy of choices were *sequentially skewed*, with *terminal behavior choices* occurring at (almost) every node. (A terminal behavior choice leads directly to the performance of some behavior, without any intervening choice necessary before the performance of that behavior.) To illustrate, consider eight behavior acts: #1, #2,...#8. If the decision three through which an individual passes in ~~time~~ (starting at the top and moving down), has the symmetrical form given in figure 5, (where first the individual decides between A1 and A2, then if A2 is chosen, between C1 and C2, etc., until he reaches the bottom level decisions, and not until then can a behavior be selected), then there is no reason to expect a Zipf curve.



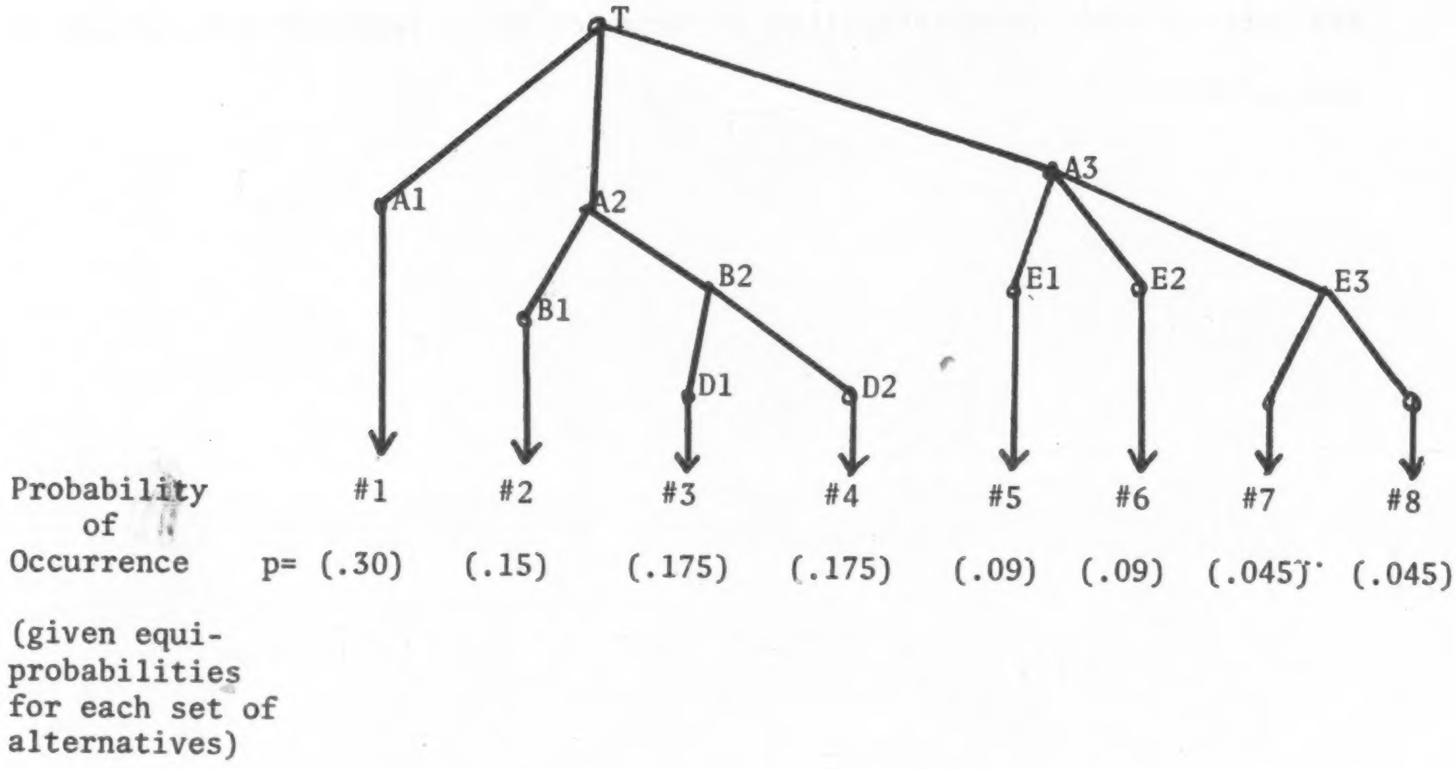
Symmetrical Decision Tree

FIGURE 5

If the probabilities are approximately equal for each alternative at each choice point, the expected rank-frequency relation would be a flat line. Sampling variability would yield a binominal distribution based on a p of .125, and the rank-frequency relation, while having some slope, would not show pronounced concavity. (The slope would have a mirror 'S' shape, but would be relatively flat.)

If the decision tree were skewed, however, so that at each node some terminal behavior choice were available, and given relatively equal probabilities at each choice point, a strongly concave curve would be obtained, although not necessarily a true Zipf curve (See Figure 6).

While it seems intuitively plausible that many decision trees would show a skewed form, no theoretical cause is apparent which would make symmetrical trees as unlikely as the empirical evidence seems to require.



Skewed Decision Tree

FIGURE 6

This discussion of decision making is extremely speculative and incompletely formulated, and is meant to serve primarily as a horizon for those who like long vistas. From this viewpoint an adequate description of an individual's behavior would be some representation of that person's decision trees, with probabilities assigned to alternative choices. Skimming off high frequency behaviors as a means of description provides only a rough approximation to such a description, but still gives a more accurate picture than characterization in terms of multi-behavior unit traits or dimensions.

BIBLIOGRAPHY

Bales, R.F., Personality and Interpersonal Behavior. New York: Holt, Rinehart and Winston, 1969.

Borgatta, E.F., L.S. Cottrell, and J.H. Mann, "The Spectrum of individual interaction characteristics: An inter-dimensional analysis." Psychological Reports, 1958, 4, 279-319.

Campbell, D. and D. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin, 1959, 56, 81-105.

D'Andrade, R.G., "Trait psychology and componential analysis." American Anthropologist, 1965, 67, #5, part 2, 215-228.

Flavell, J.H. and E.H. Flavell, "One Determinant of judged semantic and associative connection between words." Journal of Experimental Psychology, 1959, 58, 159-165.

Flavell, J.H. and D.J. Stedman, "A developmental study of judgments of semantic similarity." Journal of Genetic Psychology, 1961, 98, 279-293.

Guilford, J.P., Psychometric Methods. New York: McGraw-Hill, 1936.

Hunt, J. McV., "Traditional Personality theory in the light of recent evidence," American Scientist, 1965, 53, 80-96.

Liberman, A.M., F.S. Cooper, D.P. Shankweller, M. Studdert-Kennedy, "Percepting the Speech Code." Psychological Review, 1967, 74, 6, 431-461.

Mann, R.D., The relation between personality characteristics and individual performance in small groups. PhD. dissertation, University of Michigan, 1959.

Miller, G.A., "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review, 1956, 63, 81-97.

Miller, G.A., and E.B. Newman, "Tests of a statistical explanation of the rank-frequency relation for words in written English." American Journal of Psychology, 1958, 71, 209-218.

Mischell, W., Personality and Assessment. New York: John Wiley & Sons, 1968.

Newcomb, T., "An experiment designed to test the validity of a rating technique." Journal of Educational Psychology, 22, 279-289, 1931.

Parsons, T. and R.F. Bales, Family, socialization and interaction process. Glencoe, Ill.: Free Press, 1955.

Shweder, Richard A., "Symbolic arrangement as a measure of interpersonal behavior variables." Unpublished manuscript, 1969.

Vernon, P.E., "Ability factors and Environmental influences." American Psychologist, 1965, 20, 723-733.

Wallace, A.F.C., "On being just complicated enough." Proceedings of the National Academy of Science, 1961, 47, 4, 458-464.

A THEORY OF MARKING RULES[1]

William H. Geoghegan
University of California,
Berkeley

Introduction

In attempting to understand and describe the means by which human beings actually produce and interpret the message forms[2] appropriate to a given domain of verbal behavior, it becomes apparent that the cognitive processes involved can reasonably be characterized as information processing phenomena. Throughout the production of a message form, an individual must select and encode the information that he wants to convey (the content of his communication) and take into account those additional informational items which, though not a part of the semantic content, are nevertheless necessary to its encoding. He must gather, organize and, in general, process such information in order to identify the message form consistent with his intent and capable of an appropriate interpretation. During the interpretive process as well, an individual must consider both the message form itself and any other information necessary for an accurate decoding, and process these items in order to generate an interpretation corresponding as closely as possible to the original semantic content.

In this discussion, I want to consider a particular type of natural information processing (IP) routine that appears to be especially significant in the semantic phases of message production and interpretation.

Because of their close association with the phenomenon of semantic marking (Greenberg 1966), I shall refer to routines of this type as marking rules.[3] Following a preliminary description of the structure and operation of such routines, I will provide a detailed illustration drawn from a recent study of personal address terminology in a southern Philippine language. This example will provide the basis for some further suggestions concerning the formal properties of marking rules in general. Since work in this area is still far from complete, my comments here should be regarded as highly tentative in nature and necessarily subject to a great deal of further verification.

The Nature of Marking Rules: Preliminary Considerations

Let me begin by describing a highly simplified version of an actual information processing routine of the type that can be represented by a marking rule. Assume, first of all, that a particular individual (call him Ego) has three types of personal names that he can use in addressing people that he knows. These name-types will be termed "pet name," "nickname," and "true name," symbolized PN, NN, and TN, respectively. For each person known to Ego, assume that there is at least one lexical item corresponding to each of these three types. His daughter's true name, for example, might be "Margaret," her nickname "Peggy," and her pet name "Punkin." For each potential addressee (or "Alter"), only one of the three possible forms will represent Ego's customary, normal, or expected mode of address. For the sake of simplicity, we can assume that this form is always Alter's nickname. Thus, for example, on most occasions when Ego addresses his daughter, he uses her nickname, "Peggy," rather than her pet name or true name. Such usage may be intended only to get her attention, to single her out within a larger group of people, to direct a remark in the course of a conversation, or to perform some other function of this general nature. Beyond this, it is a perfectly neutral form of address and does not directly communicate any information about Ego's momentary attitudes or feelings toward his daughter. His use of her nickname is highly predictable, in no sense unusual, and perfectly appropriate for situations in which no attitudinal information is to be conveyed. I shall refer to this type of usage as unmarked.

Now suppose that Ego wants to convey to his daughter (or to any other addressee) a feeling of affection for her, of intimacy, happiness, etc., through the use of an address form. In this situation, he shifts from his normal (or unmarked) usage and employs her pet name, "Punkin," as the appropriate form for communicating such information. In contrast, he uses his daughter's true name ("Margaret") to indicate anger, irritation with her behavior, or some other informational item of this general nature. Such items as "affection," "intimacy," "anger," and so forth, will be referred to as marking cues (or, more simply, cues); and forms such as PN and TN will be described as marked by one or more cues with respect to the normal, or unmarked form. Thus, for example, in Ego's address system PN (pet name) is marked by "affection" with respect to NN (nickname). A marked form can be said to encode the cues by which it is marked; while an unmarked form, on the other hand, encodes none of the cues potentially available.[4]

To describe the marking rule that represents an information processing routine of this sort, I will refer to items such as PN, NN, and TN as the outputs of the rule. Each output represents a potential result of applying the routine; the identity of the output finally chosen during any particular application will depend upon the identity of input information (an unmarked output and the cues to be encoded) taken to be in effect at that time. The final output will either be unmarked (in which case no cues have been encoded), or it will be marked by one or more of the available cues.

With regard to the process of applying the routine described above, a sequence of events of the following sort takes place. The marking rule is called into play whenever Ego has to address another individual with a personal name. There is a choice between three alternative outputs (name-types), and he must decide which one will appropriately encode the information that he wants to communicate. Ego first determines which output is unmarked for the current address situation. (NN is the only possible choice in this example.) If none of the cues that can be encoded by this rule represent part of the intent of his message, then the application will be terminated; and the final output will be the unmarked form, NN. But if he wants to communicate "affection," for example, then the output currently "in effect," so to speak, would change from NN to PN through the performance of an operation appropriate to the encoding of this item. At this point, no further encoding could take place, and PN would remain in effect as the final output for this application of the routine.

I shall use the term marking operator to refer to the information processing operation that is performed when a particular cue is to be encoded, and which produces the change in effective output that represents this encoding. Each of the cues specified for a particular marking rule will be associated with a single marking operator. A given operator, however, may be associated with more than one marking cue, so long as all such cues are encoded in precisely the same fashion through all possible applications of the rule.[5]

In more formal terms, a marking operator has the properties of a many-one function defined on the set of outputs specified for a given marking rule. Suppose we have two outputs O_i and O_j for some marking rule M, and a marking operator v_k such that whenever O_i is in effect for some application of M and v_k is then applied (when one of its associated cues is encoded), the effective output changes to O_j. Assume that v_k also has the capability of encoding its associated cues when the output O_u is in effect (producing a shift to O_v), when O_x is in effect (producing a change to O_y), and so forth. The set of potential applications of v_k can be represented as the mapping

$$v_k(O_i) = O_j$$
$$v_k(O_u) = O_v$$
$$v_k(O_x) = O_y$$

etc.

With respect to the example we have been discussing, let a represent the marking operator associated with "affection," "intimacy," "happiness with Alter," etc., and let x represent the operator associated with "anger," "irritation," and so forth. The complete mapping performed by these two operators for the marking rule would then be:

$$\underline{a}(\underline{NN}) = \underline{PN}$$

$$\underline{x}(\underline{NN}) = \underline{TN}$$

A directed graph can be used as a convenient means of representing the mapping performed by each member of the set of marking operators upon the set of outputs specified for a given marking rule. Each output is represented by a single vertex of the graph, and each marking operator corresponds to one or more of the graph's arcs. If the operator $\underline{v}_k$, for example, maps the output $\underline{O}_i$ onto $\underline{O}_j$ (i.e., is capable of causing such a change in effective output), then the graph for this rule would contain an arc (labeled by $\underline{v}_k$) that is incident to both $\underline{O}_i$ and $\underline{O}_j$ and directed from the former to the latter. In Figure 1 such a graph is provided for the outputs and marking operators relevant to the example given earlier.[6]

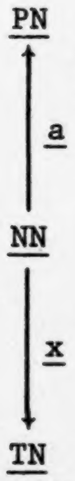


Figure 1

Input Information

The input information for a marking rule in any given application includes those items that are entered directly from external sources during the performance of the routine and which are used to determine the identity of the final output for that application. (By "external sources," I am referring to memory, perception, other information processing routines, and so forth.) In addition to the marking cues that may be entered for encoding, one of the rule's outputs must also serve as an input item. Specifically, the encoding process itself cannot be initiated until it has been decided which of the rule's outputs is appropriately unmarked for the situation in which the rule is being applied. The identity of this output must necessarily be entered as the first item of input information in any given application. When there is only one output that can take this role (as in the example we have been discussing), the selection process is completely trivial; but there exist marking rules (one of which will be described shortly) in which any one of several different outputs has the potential of being unmarked, and where the process of choosing the one appropriate to a given situation is by no means trivial. In such cases the choice of an unmarked output can determine which combinations of cues are ultimately available for encoding, and which of the rule's other outputs actually encode them.

The selection of an unmarked output will normally require the application of one or more information processing operations external to the marking rule itself. These may involve nothing more complex than retrieving and entering the necessary information from memory; but such operations may also be organized into extremely complex IP routines which themselves require an extensive amount of input information for their performance.

Once the application of a marking rule has been initiated with the selection of an unmarked output, then the options open to the user at any subsequent point in the process (so long as additional encoding operations can take place) include not only the possibility of applying an available marking operator, but also the option of using no operator at all. Once an output $\underline{O}_i$ has been identified as in effect (it may or may not be unmarked), then it would normally be possible to terminate the rule's application and leave $\underline{O}_i$ in effect as the final output. In fact, what evidence there is concerning the operation of natural marking rules suggests that the probability of utilizing any given marking operator when it becomes applicable during the encoding process is very small indeed. Observations of Samal address behavior, for example, indicate that unmarked address forms (ones that were marked by none of the cues available in several obligatorily applied rules) characterize at least 90 per cent of everyday usage. The probability of applying any individual marking operator would consequently have to fall well below the figure of 0.10.

There are marking rules that specify obligatory cues -- ones that have to be encoded when they are identified and their associated operators are applicable -- but, even so, obligatory cues seem to be in effect for such rules in only a relatively small proportion of actual instances of use. In the Samal address system, one such cue comes into effect (and must be encoded) whenever the addressee is a hajji' (an individual who has made the pilgrimage to Mecca). The proportion of actual address situations in which this occurs is quite small over the long run, since well under one per cent of the Samal population hold this particular position. Another obligatory cue used in this system must be encoded whenever the originator of the address form (Ego) does not know the addressee's name (e.g., when the output corresponding to 'nickname' is in effect and Ego does not have a 'nickname' for Alter). Since most Samal address usage is directed toward people who are well known to an individual (family members, close friends, villagemates, and so forth), this cue is encoded in only a small proportion of the situations for which the address system is actually applied.

The low a priori liklihood of a marking cue being encoded when such an operation is possible does not stem from any structural characteristic of marking rules in general. Quite to the contrary, there is some reason to believe that the existence of this phenomenon is one of the necessary conditions for the development of IP routines of the marking rule type, and that other kinds of routines normally develop when this condition does not obtain. In other words, it appears that the structural and operational characteristics of marking rules are specifically adapted to the performance of information processing tasks for which individual items of input information have very low probabilities of occurrence.[7]

If marking operators are actually applied in only a small proportion of the situations for which they could be used, this will ultimately produce a highly skewed frequency distribution for the use of a marking rule's outputs. Considering a sufficiently large number of situations in which some output $\underline{O}_i$ is unmarked, for example, $\underline{O}_i$ will remain the final output far more often than those outputs $\underline{O}_j$, $\underline{O}_k$, ..., that are marked with respect to $\underline{O}_i$; and the frequencies with which $\underline{O}_j$, $\underline{O}_k$, ..., are actually used will generally decrease in rough proportion to their degree of marking with respect to $\underline{O}_i$ (the number of operators applied when they are marked with respect to this output). This ties in with my earlier statement that unmarked usage represents "normal," or "expected" behavior for a situation, and that marked forms are in some sense "unusual" or "unexpected." To a certain extent, it is also consistent with Greenberg's (1966) use of frequency distributions as evidence for semantic marking; and it certainly accords with my own more informal estimates of the frequencies of marked and unmarked usage in Samal address.

The Name-Type Rule in Samal Address

The simple example introduced earlier in this paper allowed us to discuss the principal entities (outputs, cues, operators, etc.) involved in IP routines of the marking rule type, as well as certain properties of the information used in their application; but there are additional characteristics of structure and operation that can be illustrated only through an example of much greater complexity than the one I have been treating up to now. I would also like to move away from hypothetical cases toward a consideration of more realistic instances of this phenomenon, in order to comment on the role of marking rules within more complex semantic systems. Accordingly, I will give a brief description of the address system used by one of my Samal informants[8] and proceed from there to a more detailed discussion of one of its constituent marking rules.

The address system employed by any individual Samal is a complex information processing routine used to produce and interpret address forms. I want to stress that an address form is not a particular message (an overt act). Rather, it is a message form: a conceptual representation of an infinitely extended class of potential messages, all of which share one or more basic features in common. The cognitive representation of these features constitutes the address form in question. Those features of immediate relevance to the address system include the lexemes that are used and their temporal ordering in a given message. (I will use square brackets to set off a given address form; the plus sign will indicate concatenation and the division between lexemes.) For example, the address form

[bapa' + hajji']

consists of two lexical items, bapa' and hajji', taken in this order. This form may be realized in actual speech as either bapa' hajji' or pa' hajji' (where pa' is a common contraction for bapa'), with a wide range of possible variation in stress, intonation contour, basic pitch level, vowel length, and so forth.[9] Similarly, the address form

[dakayu' danakan]

(literally, 'one sibling') consists of a single compound lexeme -- in this case a proname (or personal name substitute).

The address system described here is capable of generating over 200 different address forms, not considering the wide range of personal names available. These can be divided, however, into ten different address form-types (AFT's), according to the classes of lexical items represented in their constituents. There are four such classes, each of which contains lexical realizations for one of the following address elements:

A: 'address term' (bapa', babu', mbo', etc.)

T: 'honorific' (tuan and dayang)

G: 'positional title' or gallal term (hajji', 'imam, maharaja, etc.)

N: 'name', including personal names ('abdul, 'ali', hasan, etc.) and pronames (name substitutes) ('oto', nde', toto', etc.)

Each of the permissible AFT's contains either one, two, or three address elements in the order: *A* > *T* > *G* > *N*. As with address forms, I will use square brackets to delimit an AFT and the plus sign to indicate concatenation. Thus, for example,

[*A* + *N*]

denotes the address form-type that contains an 'address term' and a 'name', in this order. One of its possible *lexical realizations* would be the address form

[*bapa'* + *'abdul*]

where *bapa'* is a realization for the address element *A* ('address term'), and *'abdul* is a realization for *N* ('name'). The ten permissible AFT's are:

[*A*]
[*A* + *T*]
[*A* + *T* + *G*]
[*A* + *G*]
[*A* + *N*]
[*T*]
[*T* + *G*]
[*G*]
[*G* + *N*]
[*N*]

The production of an address form requires the application of two basic groups of operations. The first comprises what I will call the

AFT-selection routine. This procedure is employed to generate the AFT whose constituent elements will be given lexical realizations in the second stage of the production process. I should mention that any AFT is capable of carrying a certain amount of semantic content *independent* of its lexical realization. Specifically, the routine employed in selecting an AFT contains two marking rules that may be used to encode the degree and type of 'respect' (*'addat*) that Ego wants to convey to the addressee.

The second group of operations, referred to here as the *lexical realization phase* of production, consists of five separate information processing routines. There is one lexical realization routine for each of the address elements *A*, *T*, and *G*, and two such routines for the element *N*. Of the latter, one of the two is used when *N* appears in an AFT that also includes either *A* or *G* (there is no AFT containing both *T* and *N*); the other is applied when the AFT [*N*] has been chosen. In the former case, *N* can be realized only by a personal name (or by Ø, if a name is not known). When *N* is the *only* constituent of the AFT, however, it may be realized by either a personal name or by a proname. It is the IP routine used in the latter situation (the *name-selection routine*) and the marking rule used to perform one of its major operations that will be of particular concern to us in this discussion.

When an address form of type [*N*] is to be used, the selection of a realization for *N* proceeds in two phases. In the first of these, a marking rule (the *name-type rule*) is used to determine the *type* of personal name or proname that appropriately encodes the information (in addition to that carried by the AFT) that Ego wants to convey to the addressee. Once this decision has been made, a second operation must be performed in order to determine which lexical item is the proper realization for the name-type just selected.

There are seven types of personal names and pronames recognized by this informant (and other adult Samal speakers), each of which corresponds to an output of the name-type marking rule. These name-types are listed below, with brief descriptions and examples for each.

TN: *'ōn-na to'od* 'his true name' (The personal name commonly recognized as an individual's actual, "given" name; e.g., *'abdulmuluk*.)

TN': *'ōn-na to'od* 'his true name'[10] (Proname-type corresponding to *TN*: e.g., *lella* ['man'/'male'], *matto'a* ['old person'], etc.)

NN: *danglay-na* 'his nickname' (A personal name frequently derived from an individual's 'true name' [e.g., *'abdulmuluk* → *muluk*]; often the name by which he is referred to in the community at large.)

NN': *danglay-na* 'his nickname' (The proname-type corresponding to *NN*: e.g., *dakayu' siali* ['one younger-sibling'].)

PN: *'ugay'ugay-na* 'his pet name' (Also, *'ugay'ugay-ku ma 'ia* ['my pet name for him']. A personal name frequently derived from an individual's 'nickname' [e.g., *muluk* → *llu'*, *hakim* → *kki'*]. 'Pet names' are highly idiosyncratic, and may be derived from a number of sources other than 'nicknames': i.e., physical characteristics [e.g., *sombeng* ('harelip') → *'ombeng*], past events, nonsense words, etc.)

PN': *'ugay'ugay-na* 'his pet name' (The proname-type corresponding to *PN*: e.g., *'oto'*, *nde'*, etc.)

T: This is a special proname-type that does not correspond directly to any type of personal name, but is rather derived from the Samal 'honorific'. It conveys a high degree of 'affect', and takes the realizations *tuan* (for males) and *dayang* (for females).

The personal name-types include *TN*, *NN*, and *PN*. Realizations for these are names of specific individuals and are determined in any particular address situation by the identity of the individual being addressed. The proname-types include three (*TN'*, *NN'*, and *PN'*) whose realizations may serve under certain conditions as direct substitutes for personal names of the corresponding types (*TN*, *NN*, and *PN*, respectively). Lexical realizations for these are chosen by applying a code rule (Geoghegan 1968, 1970) which requires consideration of the addressee's age group, sex, stage of development (for children), and relative age (for addressees of Ego's age group). Realizations for *T* (the fourth type of proname) are chosen according to the addressee's sex.

The marking rule used to select a name-type involves five different marking operators, four of which are optional in use and are associated with cues representing information concerned with certain attitudes and emotional states ('affect', 'anger', etc.). One operator (*n*) is obligatory when an associated cue characterizes the address situation. The five operators and their associated marking cues can be listed as follows:

a: 'positive affect':

'alasa (*'aku*) *ma* *'ia* '(I) like him'/'(I) feel affectionate towards him'

kinōgan *'atay-ku* *ma* *'ia* 'my liver is made happy for him'

a': 'negative affect':

ngga'i (*'aku*) *'alasa* *ma* *'ia* '(I) do not like him'/'(I) do not feel affectionate towards him'

'ala'at *'atay-ku* *ma* *'ia* 'my liver is bad for him'

'akuddu' *'atay-ku* *ma* *'ia* 'my liver is upset/disturbed by him'

x: 'anger':

mag'amā *'aku* *ma* *'ia* 'I am angry at him'

nidugalan *'aku* *ma* *'ia* 'I am made upset/nauseated by him'

d: 'deference':

magmaltabat *'aku* *ma* *'ia* 'I defer to (show mild respect for) him' n: 'Alter (Alter's name) not known':

'insa' 'ia kinata'uhan ku 'he is not known to me'

ngga'i kata'uhan ku ma 'ia 'I do not know him'

ngga'i kata'uhan ku X-na 'I do not know his X (name-type)'

(This list contains the descriptive phrases used by the informant to characterize the information that can be communicated through selection of a name-type.) I should point out that the cues associated with any one of these operators are essentially synonymous with one another, insofar as they connote particular attitudes or emotions; and they are, in fact, used interchangably by informants in statements regarding the use of personal address. (This may not be altogether clear from the rather literal translations of the Samal descriptive phrases.) For this reason, and to help simplify matters somewhat, I will refer to the cues associated with a given operator by a single collective gloss that stands for and roughly characterizes the information involved. Thus, for example, the gloss 'negative affect' will be used for the three cues associated with the marking operator a'. The mappings performed by these operators upon the outputs of the name-type rule are shown in the directed graph of Figure 2.

I stated earlier that a marking rule could have more than one unmarked output. This version of the name-type rule is a case in point, since it allows for either TN, NN, or PN to be used in this manner.[11] In any given application, the selection of an unmarked output is based on Ego's 'habitual' address usage to Alter (kabiaksahan pangōn, 'habitual means-of-naming'). If there has been a past history of interaction with the addressee sufficient for the establishment of a 'customary' or 'habitual'

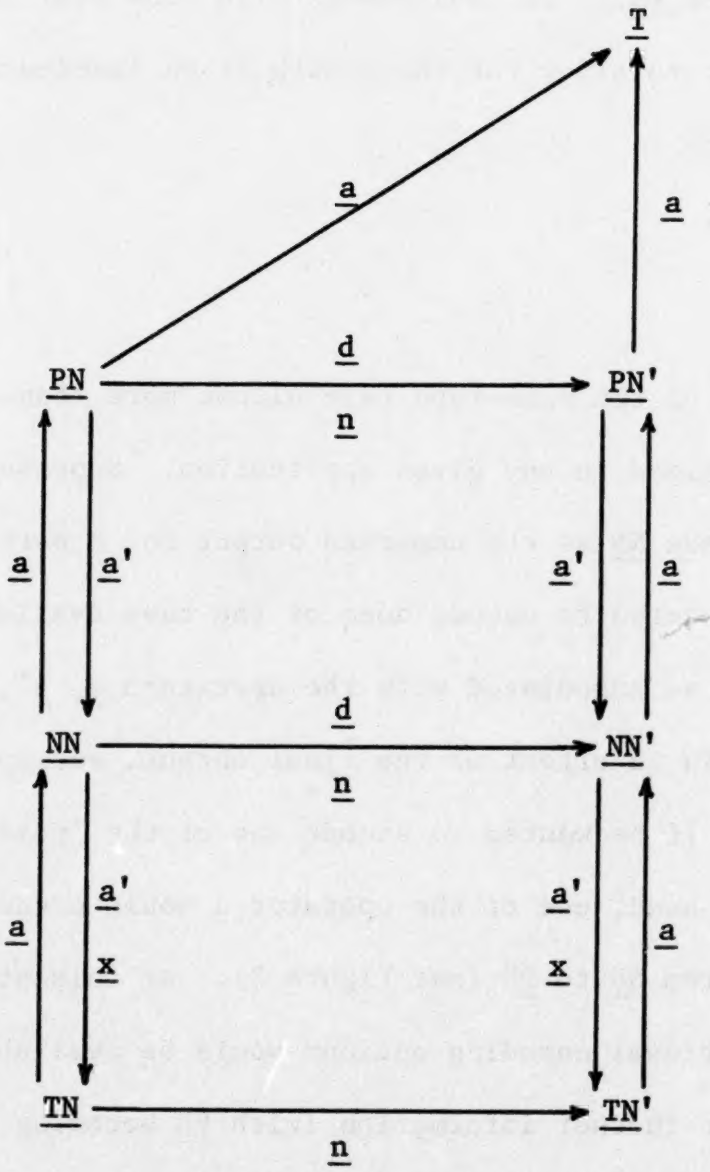


Figure 2

name, then the name-type corresponding to this form will be taken as unmarked. (Such an 'habitual name' is always interpreted as a realization of either TN, NN, or PN.) For addressees with whom past interaction has not been sufficient to allow for the growth of an 'habitual name', NN is taken as unmarked.[12]

Marking Sequences

The structure of the name-type rule allows more than one marking operator to be utilized in any given application. Suppose, for example, that Ego were to take NN as the unmarked output for a particular address situation. If he wanted to encode none of the cues available at this point in the process (those associated with the operators a, a', n, d, and x), then NN would remain in effect as the final output, and application of the rule would cease. If he wanted to encode one of the 'positive affect' cues, on the other hand, use of the operator a would occasion a shift in effective output from NN to PN (see Figure 2). At this stage of the process, several additional encoding options would be available. Ego could choose to encode no further information (with PN becoming the rule's final output), or he could continue the application by encoding information associated with either n, d, or a.[13] (For reasons to be discussed shortly, a' would not be applicable once 'positive affect' had been encoded.) If Ego did not have a 'pet name' for Alter, then n would have to be applied (it is obligatory in such situations), producing a shift from PN to PN'. The same change in effective output would occur if he encoded 'deference' (d).[14] Once again, Ego would have the option of terminating his application of the rule, or of continuing with further encoding operations. The only available operator at this point in the process is a ('positive affect'), the use of which would produce a shift from PN' to T.

Another series of encoding operations (when NN is unmarked) might involve the initial use of n (if Alter's 'nickname' were unknown to Ego), causing a shift from NN to NN', followed by the encoding of 'anger' (x). The latter operation would produce a change in effective output from NN' to TN', at which point application of the marking rule would have to cease, leaving TN' in effect as the final output. (Although it appears that a is applicable at this point, 'positive affect' cannot be encoded simultaneously with 'anger'.) In general, a particular sequence of encoding operations (a series of successively applied marking operators) is possible only if we can discover in the graphical representation of the rule a path progression[15] corresponding to this sequence and originating with the effective unmarked output. In Figure 2, for example, we can find the path progression $\langle a, d, a\rangle$ originating at NN and terminating at T; and we also have the progression $\langle n, x\rangle$ originating at NN and terminating at TN'.

Given the set O of outputs for some marking rule M and a set V of marking operators v_i for M (where each such v_i is a many-one function defined on O), I will represent such a series of operations by a *marking sequence* V_k defined on O and V. In formal terms, a marking sequence V_k can be defined as an ordered n-tuple of marking operators v_i, where

$$V_k = \langle v_1, v_2, \ldots, v_{n-1}, v_n\rangle,$$

and for which the order of operators in V_k corresponds directly to the order of encoding operations in some potential application of M. Although an ordered n-tuple is normally taken to include at least two elements, we shall find it helpful to modify this practice in the representation of marking sequences. Specifically, we should allow for sequences that include only one operator (to represent applications for which only one marking operation occurs); and we shall also find it useful to allow for the limiting case of an "empty sequence" which contains no marking operators. The latter will be denoted

$$\underline{V}_0 = \phi$$

and will be used to characterize those applications of a marking rule in which no cues are encoded and for which the unmarked and final outputs are identical. Two of the marking sequences possible for the name-type rule when NN is unmarked are the ones described earlier: namely,

<a, d, a> and
<n, x>.

Permissible Applications

Given the set O of outputs and the set V of marking operators upon which a particular rule is constructed, it may be possible to define one or more marking sequences that correspond only to impermissible applications of the IP routine in question. An "impermissible application" is one that, while formally possible under the specification of O and V, encodes a body of information that would be rejected by native speakers as "meaningless," self-contradictory, or in some other way semantically inappropriate. This will normally occur when the marking sequence representing a potential application contains the elements of a *cycle* (described below) or, in general, any pair of marking operators associated with semantically incompatible cues.

Referring to Figure 2, we can see that there are several potential marking sequences that correspond to path progressions beginning and terminating at the same output. The sequence <*a*, *a*'>, for example, leads from *NN* to *PN*, and then back to *NN*. (It also corresponds to a series of arcs beginning and ending at *TN*.) Similarly, the potential marking sequence <*x*, *a*> would, if *NN* were the unmarked output, simply map this output onto itself. I will refer to any marking sequence of this type as a *cycle*. When we find two or more marking operators that form a cycle, it will generally be the case that their associated cues are in some way mutually contradictory. That is, if a cue associated with one of these operators is in effect for a given situation, then cues associated with one or more of the other operators in the cycle cannot reasonably be in effect for that situation. All of the cycles that can be defined for the name-type rule contain either *a* and *a*', or *a* and *x*. Some of the cues associated with *a* and *a*' are in direct "logical" conflict with one another (e.g., *'alasa* *ma* *'ia* ['feel affectionate towards him'] and *ngga'i* *'alasa* *ma* *'ia* ['do not feel affectionate towards him']), while others associated with these two operators conflict through their conceptual similarity (essentially synonymy) to "logically" incompatible cues. With regard to those associated with *a* and *x*, the conflict is of a more indirect nature. 'Anger' (*'amā*) and a sort of 'sick rage' (*dugal*), characterizing the cues associated with *x*, are normally taken to imply the concurrent existence of 'negative affect' (*a*'); which, in turn, implies the absence of 'positive affect' (*a*). As momentary attitudes or emotions, 'anger' and 'positive affect' are thus regarded as incompatible with one another.

A close inspection of Figure 2 should also indicate that even if the operators forming a cycle were utilized in the application of this rule, they would "neutralize" each other in terms of their effect on the identity of the final output. If *NN* were unmarked, for example, and if use of the operator *a* were followed by the use of *a*', then *NN* would be the final output of the rule -- that is, an output unmarked by cues associated with either of the two operators. Similarly, suppose that the marking sequence <*a*, *d*, *a*'> (which contains all elements of the cycle <*a*, *a*'>) were to be applied when *NN* was unmarked. The final output in this case would be *NN*', an output that is marked only by 'deference' (*d*) with respect to *NN* (assuming that *n* is not in effect). The cues associated with *a* and *a*' would be "cancelled out," so to speak, leaving 'deference' as the only informational item belonging to a reasonable interpretation for the use of *NN*' in this situation. In brief, even if Ego attempted to use this marking rule to encode cues associated with the elements of a cycle, these items of information could not be communicated by a single name-type because they would be "neutralized" during the encoding process itself.[16] (This phenomenon appears throughout those versions of the Samal address system that have been analyzed in detail.)

In addition to marking sequences of the type just described, there are others which, for one reason or another, represent impermissible applications of a given marking rule. Once again, such sequences generally contain two or more marking operators associated with contradictory or conflicting cues. Such conflicts are held to exist, for example, between 'deference' and 'negative affect', and between 'deference' and 'anger', for the Samal name-type rule. This follows from the fact that an overt display of 'deference' normally connotes a certain degree of 'respect' ('addat) on Ego's part, while the overt expression of 'negative affect' or 'anger' implies a definite lack of 'respect'. On several occasions, informants have stated explicitly that the encoding of such conflicting information is incorrect; and attempts to elicit these cues in the interpretation for such usage (e.g., the use of NN' when PN is the habitual or unmarked form) have consistently failed. (When PN is unmarked, the use of NN' is taken by informants to imply 'negative affect' and the fact that Ego has forgotten Alter's 'pet name' or 'nickname', an eventuality that would lead to the use of n rather than d.) Accordingly, it would seem appropriate to eliminate from the set of permissible marking sequences for a given rule all those which contain operators associated with such conflicting cues. In many respects, this restriction is identical to the one concerning sequences that contain the elements of a cycle (which also involve conflicting cues). The major difference is that cycles can be discovered on purely structural grounds (i.e., from the mappings performed by the various marking operators); while in the present case, structural criteria would normally be absent. (There is nothing about the mappings performed by d and a', for example, that would lead us to conclude that their associated cues are incompatible with one another.)

While certain marking sequences should be ruled out as characterizing impermissible or inappropriate applications of a marking rule, there are other sequences of operators that should definitely be included as representing necessarily permissible applications. Referring once again to Figure 2, suppose we are given the marking sequence <a, a, d> as representing one appropriate use of the name-type rule when *TN* is the unmarked output. Assuming that the operator *d* is not made obligatory by the previous encoding of 'positive affect' (either once or twice), nor by the fact that *TN* is unmarked, then it should be possible to find applications of the rule for which it was permissible *not* to encode 'deference', and for which the marking sequence <*a*, *a*> represented the operations actually performed. Similarly, it should be possible to find applications for which it was permissible to encode 'positive affect' only once, and other applications for which no marking cues need be encoded and for which *TN* remained as the final output. In other words, given that <*a*, *a*, *d*> represents a permissible application when *TN* is unmarked, then each of the following marking sequences must also represent a permissible application under such conditions:

<*a*, *a*>
<*a*>
ϕ

(Recall that ϕ represents a "sequence" of operations in which none of the available marking operators is applied, and in which only an unmarked output is chosen.)

Phrasing the above in more general terms, if V_i is a marking sequence that represents a permissible application when O_u is unmarked, then each marking sequence constructed on V_i by taking the first *m* elements in order (where V_i contains *n* elements, and $n \geq m \geq 1$) also represents a permissible application of the rule when O_u is unmarked. Available evidence concerning the structure and operation of naturally occurring marking rules suggests that this should be taken as a characteristic of such information processing routines in general.

Table 1 contains a list of the thirty-five marking sequences that characterize permissible applications of this informant's version of the name-type rule. Each of these may be applied in conjunction with any of the three potentially unmarked outputs (*PN*, *NN*, and *TN*) for which the formal possibility exists. The sequence of operations represented by <*a*, *n*> (V_9), for example, may be applied in situations for which either *TN* or *NN* is unmarked. (In Figure 2, there is a corresponding path progression originating from each of these outputs.) This sequence may *not* be applied when *PN* is unmarked, because *a* maps *PN* onto *T*, and *T* does not fall within the domain of *n*. (That is, the operator *n* cannot be applied when *T* is the effective output. Figure 2 will show that there is no path progression corresponding to <*a*, *n*> originating at *PN*.)

The list shown in Table 1 was gathered through intensive elicitation sessions with the informant whose version of the name-type rule has been presented here. It should be noted that none of the marking sequences in this list are cycles, nor do any of them contain all the elements of a cycle. Similarly, there is no marking sequence in this set that contains both *d* ('deference') and *a*' ('negative affect') or both *d* and *x* ('anger'). It should also be noted that for every marking sequence V_i in this list, such that V_i contains *n* operators ($n \geq 1$), there is another sequence V_j in this set containing only the first *m* elements of V_i ($n \geq m \geq 0$) in the same order. With respect to the sequence V_{30}, for example, we have:

Table 1

$\underline{V}_0 = \phi$

$\underline{V}_1 = \langle \underline{a} \rangle$

$\underline{V}_2 = \langle \underline{a}, \underline{a} \rangle$

$\underline{V}_3 = \langle \underline{a}, \underline{a}, \underline{a} \rangle$

$\underline{V}_4 = \langle \underline{a}, \underline{d} \rangle$

$\underline{V}_5 = \langle \underline{a}, \underline{a}, \underline{d} \rangle$

$\underline{V}_6 = \langle \underline{a}, \underline{a}, \underline{d}, \underline{a} \rangle$

$\underline{V}_7 = \langle \underline{a}, \underline{d}, \underline{a} \rangle$

$\underline{V}_8 = \langle \underline{a}, \underline{d}, \underline{a}, \underline{a} \rangle$

$\underline{V}_9 = \langle \underline{a}, \underline{n} \rangle$

$\underline{V}_{10} = \langle \underline{a}, \underline{a}, \underline{n} \rangle$

$\underline{V}_{11} = \langle \underline{a}, \underline{a}, \underline{n}, \underline{a} \rangle$

$\underline{V}_{12} = \langle \underline{a}, \underline{n}, \underline{a} \rangle$

$\underline{V}_{13} = \langle \underline{a}, \underline{n}, \underline{a}, \underline{a} \rangle$

$\underline{V}_{14} = \langle \underline{a}' \rangle$

$\underline{V}_{15} = \langle \underline{a}', \underline{a}' \rangle$

$\underline{V}_{16} = \langle \underline{a}', \underline{x} \rangle$

$\underline{V}_{17} = \langle \underline{a}', \underline{x}, \underline{n} \rangle$

$\underline{V}_{18} = \langle \underline{a}', \underline{n}, \underline{x} \rangle$

$\underline{V}_{19} = \langle \underline{a}', \underline{n} \rangle$

$\underline{V}_{20} = \langle \underline{a}', \underline{n}, \underline{a}' \rangle$

$\underline{V}_{21} = \langle \underline{a}', \underline{a}', \underline{n} \rangle$

$\underline{V}_{22} = \langle \underline{x} \rangle$

$\underline{V}_{23} = \langle \underline{x}, \underline{n} \rangle$

$\underline{V}_{24} = \langle \underline{n} \rangle$

$\underline{V}_{25} = \langle \underline{n}, \underline{a} \rangle$

$\underline{V}_{26} = \langle \underline{n}, \underline{a}, \underline{a} \rangle$

$\underline{V}_{27} = \langle \underline{n}, \underline{a}, \underline{a}, \underline{a} \rangle$

$\underline{V}_{28} = \langle \underline{n}, \underline{a}' \rangle$

$\underline{V}_{29} = \langle \underline{n}, \underline{a}', \underline{a}' \rangle$

$\underline{V}_{30} = \langle \underline{n}, \underline{a}', \underline{x} \rangle$

$\underline{V}_{31} = \langle \underline{n}, \underline{x} \rangle$

$\underline{V}_{32} = \langle \underline{d} \rangle$

$\underline{V}_{33} = \langle \underline{d}, \underline{a} \rangle$

$\underline{V}_{34} = \langle \underline{d}, \underline{a}, \underline{a} \rangle$

$$\underline{V}_{30} = \langle \underline{n}, \underline{a}', \underline{x} \rangle$$

$$\underline{V}_{28} = \langle \underline{n}, \underline{a}' \rangle$$

$$\underline{V}_{24} = \langle \underline{n} \rangle$$

$$\underline{V}_{0} = \phi$$

In other words, the set of permissible applications of this rule is consistent with the various constraints on such applications discussed in this section.

The Process of Interpretation

The existence of more than one potentially unmarked output for the name-type rule raises an important point concerning its use in both the production and interpretation of address forms. Specifically, neither the encoding nor the decoding process can take place without prior identification of the unmarked output in effect for a given application of the rule. This stems from the fact that the correspondence between an output and the information that it encodes varies according to the identity of the unmarked output chosen for a given application. Consider the various encodings shown in Table 2. If 'positive affect', for example, is the only item of information

Table 2

	Unmarked Output:			
Cues Encoded:	TN:	NN:	PN:	
'positive affect' (<a>)	NN	PN	T	Final Output
no cues encoded (ϕ)	TN	NN	PN	
'negative affect' (<a'>)	*	TN	NN	

*** The operator a' cannot be applied when TN is in effect.**

to be communicated, then it is appropriately encoded by a 'nickname' when TN is unmarked, by a 'pet name' when NN is unmarked, and by a proname of type T when PN is the unmarked output. Without prior identification of the unmarked form, no decision could be made regarding the proper method of encoding this item.

With respect to the interpretive process, suppose that Ego addressed Alter with the latter's 'nickname'. If Alter were unaware of the output that Ego originally took as unmarked, then he would have no way of determining which of three conflicting interpretations was the correct one. More precisely, a 'nickname' is marked by 'positive affect' when TN is unmarked; it is marked by 'negative affect' when PN is unmarked; and it encodes no cues at all when NN itself is the unmarked output. Not only do the three possible interpretations differ, but two of them ('positive affect' and 'negative affect') are in complete conflict with one another. If Alter could decide which output Ego took as unmarked, then he would be able to provide the message with a more or less unambiguous interpretation; but should he disagree with Ego in this identification, then misinterpretation would be the inevitable result.

In other words, for this marking rule to be at all effective in communicating information, there must be some procedure for identifying the unmarked output each time the rule is applied, during both production and interpretation; and this procedure must normally lead to agreement between Ego and Alter. During the production process, as I mentioned earlier, a rather brief series of operations is employed for this purpose. Ego would first attempt to identify his 'habitual name' (kabiaksahan pangōn) for Alter.

If such a name existed, then it would be entered as an item of input information to the address system, and would be analyzed to determine its corresponding name-type. The result would then be entered as the unmarked output for this application. If there were no 'habitual name', on the other hand, then NN would be taken as unmarked.

This series of operations is also applied, however, during the initial phases of the interpretive process; even though it is basically a production routine. When employed in this manner, the routine is applied as it is assumed to have been used during production of the address form: that is, from Ego's point of view. Alter would first attempt to identify Ego's 'habitual name' for him; and, given that this can be done, determine the corresponding name-type to be entered as the unmarked output. If such a name could not be identified, then Alter would take NN as unmarked. Following the performance of these operations, the decoding of cues could take place. Suppose Alter had been addressed by a 'pet name', and that Ego's customary address usage to Alter was the latter's 'nickname'. In the initial phase of interpretation (what I shall refer to from now on as the production phase), Alter would identify NN as the output Ego probably took as unmarked. He would also analyze the address form actually used and note that it contained a realization for PN (taken to be the final output). Comparison of these two outputs within the marking rule would indicate that 'positive affect' had been encoded; and this would be the interpretation assigned to Ego's original message. Without identification of the unmarked output, the address form would have had an ambiguous interpretation, since a name of type PN could encode either no cues (when PN is unmarked), 'positive affect' (when NN is unmarked), or "extreme" 'positive affect' (when TN is unmarked and the operator a is applied twice in succession).

While the production phase of the interpretive process involves a relatively simple IP routine for the name-type rule, in other segments of the Samal address system the production phase can become quite complex. Selection of an address form-type (AFT), for example, requires that one of two "AFT marking rules" be applied. (These rules are used to encode information concerning the degree and type of 'respect' ['addat] that Ego wants to communicate to Alter.) Determining the unmarked output for either one can involve the use of two additional IP rules (a code rule and a marking rule) as well as a number of subsidiary operations for the input and analysis of necessary information. Items concerning Ego's 'habitual' address form for Alter, Alter's age group relative to Ego's, Alter's status as a hajji' (one who has made the pilgrimage to Mecca), and so forth, may all become relevant to the choice of an unmarked AFT.

Although items such as these ('habitual' forms, relative age, and so forth) are employed in the production process and affect the identity of the final address form, they cannot realistically be described as representing any portion of the primary intent of such a message. That is, they are not items that Ego would normally encode in an address form simply for the purpose of communicating them to Alter. Alter can reasonably be expected to know whether or not Ego has an 'habitual' address form and name for him, the actual identity of such forms, his age relative to Ego's, his own status as a hajji', and (with regard to other portions of the address system) his sex, age group, kin relation to Ego, and so forth. Rather, such information has the primary role of establishing a framework or context in which other, more immediately salient items of information can be encoded or decoded. Consider the nature of information used in the name-type rule. It includes such items as 'affection' for Alter, 'happiness' with him, attitudes of 'deference', 'dislike', 'anger', and so forth. These are concerned with relatively private emotional and attitudinal states; and they are normally encoded in an address form only when Ego explicitly wants such information to have an effect on Alter's current state of knowledge, on his attitudes, and on the behavior that might be expected to follow from such. When actually encoded in an address form, items of this type exemplify what I would prefer to call the semantic content of the message -- information that Ego is deliberately attempting to convey to a particular individual.

I should emphasize, however, that by far the largest proportion of actual address usage in Samal is completely unmarked, and devoid of semantic content in the present sense of the term. (Most instances of personal address usage have what appears to be a metacommunicative function: they signal the opening of communication, direct messages to specific individuals, serve to emphasize portions of a complex utterance, and so forth.) For the relatively small proportion of cases in which content information is actually encoded in an address form, this is invariably done through the use of marking rules, a fact that holds for every version of the address system elicited during the course of this study.[17]

It is particularly important to maintain a strong distinction between content and context information (where the latter includes items such as an 'habitual name', relative age, etc.) if we are to understand adequately this type of communicative process. These two types of information differ not only in regard to communicative intent, but they play entirely different roles during the production and interpretation of address forms. This is best illustrated by the interpretive process, where context items are used during the production phase (to "generate" unmarked outputs for the various marking rules) in much the same manner as during the actual production of an address form, and where content items represent information retrieved from a message form during the decoding phase of interpretation. The fact that context information is normally shared between the parties in an address situation is from this standpoint not accidental, but a necessary precondition to effective communication. Given that content information in the address system is encoded and decoded through the use of marking rules, and that Ego and Alter must agree on the identity of the unmarked output in effect for a given rule during any address situation, it follows that there must also be agreement between the two individuals on the identity of those context items that are used in selecting the unmarked output. If such items were not shared, then disagreement on the identity of the unmarked output would be likely to follow; and this, in turn, could easily lead to miscommunication (where the information encoded by Ego does not correspond to that decoded by Alter).

In summary, an adequate understanding of the Samal address system -- as a device for interpersonal communication -- cannot be obtained unless we make a careful distinction between those items of information that may be used as part of a message's content and those which are used to establish a context for the encoding and decoding of content items. It is important to note, also, that the bulk of the information normally used in the production of an address form is of the context variety, is necessarily shared between the two parties to a communication in normal address usage, and does not represent a portion of the sender's communicative intent. As I suggested earlier, this implies that most instances of personal address usage involve forms that are relatively free of semantic content (that is, which are used to transmit little or no internally encoded information). When content information is encoded during an application of the address system, however, this is invariably done through the use of one or more marking rules.

FOOTNOTES

1. The data upon which this paper is based were collected during a study of personal address terminology among the Balangingi' Samal (a Muslim group of the southern Philippines). Fieldwork in the Philippines was supported by the National Institute of Mental Health (research grant number MH-). During subsequent periods of research and analysis, support was generously provided by the Institute of International Studies and the Committee on Research (both of U.C., Berkeley), the Office of Education (Department of HEW: grant number OEG-9-9-140281-0038(057), John Gumperz principal investigator), and by an NIMH grant (USPHS MH-18188-01: Paul Kay principal investigator) for the Language-Behavior Research Laboratory at Berkeley. The assistance of these agencies is gratefully acknowledged. Brent Berlin, Roy D'Andrade, Charles Frake, Paul Kay, and Robert Randall have provided many helpful comments and criticisms during informal discussions; and while this paper has benefitted greatly from their assistance, its errors and omissions are solely my own.

2. For purposes of this discussion, a message form will be regarded as a cognitively localized configuration of information that may be realized by, or represent, any one of a number of alternative, though equivalent messages. The message itself is an overt act that may vary on a number of attributes not directly relevant to the identity of the message form which it realizes (e.g., in terms of certain paralinguistic features).

3. A preliminary, and highly informal discussion of this topic was presented in an earlier paper (Geoghegan 1969). For a more general treatment of information processing systems and rules, and a tentative formalization for a theory of marking rules, see Geoghegan (1970).

4. The actual use of an unmarked form does not imply that the attitudes or feelings represented by available cues do not currently exist, but only that Ego has not chosen to communicate such information through his choice of a name-type. Moreover, if Ego had decided to communicate "anger" by addressing Alter with the latter's true name, this would not necessarily imply that the attitude was true in some objective sense, but rather that Ego had simply chosen, for whatever reason, to communicate such information to the addressee. (In reprimanding his daughter, for example, Ego might address her as "Margaret" [TN] in order to communicate "anger," even though actually amused by her misbehavior.) What we are concerned with here is the process by which an individual goes about encoding information once he has decided to communicate it to another individual -- not with the truth or falsity of what he wants to say.

5. In naturally occurring IP routines of the marking rule type, it generally appears to be the case that cues which are encoded in the same manner are conceptually, or semantically, quite similar to one another. In an example presented later in this paper, there are several cues that in native usage correspond to minor variations of a more basic concept that might be labeled 'positive affect'. Each of these cues is encoded in precisely the same way, and they are regarded by informants as essentially synonymous in personal address.

6. I should mention that such a graph does not normally contain all the information necessary for the complete specification of a given marking rule. It describes only the mapping performed by each of the rule's operators, but it does not indicate the range of permissible applications of the rule itself.

7. A detailed discussion of this point would take us far beyond the limited scope of this paper. For a more complete treatment of the acquisition and development of information processing rules and a discussion of possible cognitive mechanisms relevant to this process, see Geoghegan (1970).

8. I conducted a study of the semantics of personal address among the Balangingi' Samal in 1966-67 in the *barrio* of Tagtabon (Tictauan Island), approximately six miles east of Zamboanga City center (Mindanao). Nine informants were interviewed at length on this subject, and the address systems of six were selected for detailed analysis. The present discussion concerns the system used by a woman who was 56 years old at the time of the study. While there is extensive variation between informants as to the details of personal address, the basic structure of the system is the same in all cases; and each of the informants utilized a marking rule, similar in most respects to the one given here, for the selection of name-types.

9. These paralinguistic features can be semantically important, but their selection depends on the use of IP routines that are relevant to verbal behavior in general and which lie outside the address system proper.

10. The name-types *TN'*, *NN'*, and *PN'* are sometimes referred to by an expression such as:

X *'ia* *'ōn-na* *to'od* *bang* *ngga'i* *kata'uhan* *ku* *'ōn-na* (*to'od*).
"*X* is his 'true name' when I do not know his (true) name."

11. Other versions of the rule (elicited from different informants) may vary on this point. Name-type rules used by children, adolescents, and young adults usually allow for only one unmarked output (*NN*). Rules used by older adults permit two or three outputs to be unmarked. (The address system, and especially the name-type rule, continues to develop in the direction of greater complexity until an individual is about 50 or 60 years old.)

12. The name by which an individual is normally known and referred to in a community is usually described as his 'nickname' (*danglay*). Although Ego's 'habitual' usage to a given addressee may be a 'pet name' or 'true name', this form would not generally be used in reference, except to persons who customarily address the individual in this manner.

13. In using the name-type rule, it is possible to encode either 'positive affect' (*a*) or 'negative affect' (*a*') more than once in a single application. This has the effect of communicating a more intense or stronger version of the attitude in question. In situations for which *TN* is unmarked, for example, 'positive affect' could be encoded once (with *NN* as the final output if no other cues were relevant), twice (with *PN* as the final output),

or three times (with *T* as the final output). The use of *T* in an address form would communicate a higher "degree" of 'positive affect' than would *PN*; and *PN* would indicate a more intense attitude than *ŃN*.

14. 'Positive affect' and 'deference' represent a fairly frequent combination of marking cues in Samal usage. The act of communicating these items (through address or otherwise) is referred to by the term *bijjak* ("to cajole") and is often performed by an individual when he desires the recipient of his message to do him a large favor.

15. A path progression is composed of a sequence of arcs, each consecutive pair of which contains arcs adjacent to a single vertex (output) such that one arc is incident into the vertex and the other is incident out of the vertex. (See Busacker and Saaty 1965:27.)

16. This is not meant to imply that the Samal are incapable of irony in personal address (i.e., the encoding of contradictory attitudinal or emotional cues), but rather that irony cannot be effectively communicated through the use of a single marking rule. One can convey 'positive affect' and 'negative affect' simultaneously, for example, by applying the operator *a* in the name-type rule, and by applying an operator encoding 'negative affect' in one of the marking rules used to generate paralinguistic features of an utterance.

17. I suspect that this might be true for address systems in general. It certainly holds for the Samal system, and seems to be the case for the American English and Bisayan systems as well. There is no *a priori* reason why marking rules should be the *only* type of IP routine capable of encoding content items. This phenomenon seems rather to stem from the relatively infrequent use of marked address forms (quite likely universal to such systems), and the fact that IP routines of the marking rule type are particularly well adapted to the efficient encoding of infrequently used items of information. (See Geoghegan 1970.)

REFERENCES CITED

Busacker, Robert G., and Thomas L. Saaty

1965 Finite graphs and networks: an introduction with applications. New York, McGraw-Hill.

Geoghegan, William H.

1968 Information processing systems in culture. In Paul Kay, ed., Explorations in mathematical anthropology. Cambridge, MIT Press. (In press.)

1969 The use of marking rules in semantic systems. Working Paper No. 26. Language-Behavior Research Laboratory. University of California, Berkeley.

1970 Natural information processing rules: formal theory and applications to ethnography. Unpublished Ph.D. dissertation. Anthropology Department, Stanford University.

Greenberg, Joseph H.

1966 Language universals. The Hague, Mouton & Co.

Utilization of Cognitive Structures in Problem Solving and Reasoning

James G. Greeno
The University of Michigan

This paper was prepared for a Workshop on Cognitive Organization and Psychological Processes sponsored by the Committee on Basic Research in Education, National Research Council.

The research reported here was supported by Grant OEG-0-9-320447-4194 (010) from the U.S. Office of Education, Basic Research Branch, and by Contract AF 49(638)1736 between the Advanced Research Projects Agency and the Human Performance Center of the Department of Psychology, University of Michigan.

In this paper I will be concerned with problem solving of a rather routine variety. I take it that in most problem-solving situations people use cognitive structures that they have acquired in the past and that they retrieve from memory in order to solve the problem at hand. A clear case of this kind of problem solving is the solution of a simple word problem, using a simple formula of some kind. You or I would be in such a situation if we were asked to calculate the standard score corresponding to a score of 88 given that the mean score was 64 and the standard deviation was 16.

In Fig. 1 I have drawn a flow chart representing an innocent theory of this kind of problem solving. The boxes are shown as overlapping rather than being strung out in the usual way to indicate the probable fact that the functions represented there do not occur in a strictly serial fashion. I suppose there would be little argument against the idea that something like Fig. 1 represents the process of solving a simple verbal problem that involves a formula. The question, then, is how to obtain information that could lead toward a theory of this processing and be closer to an acceptable degree of specificity.

The work that I will describe first has been directed toward the understanding of processes in the third and fourth levels of the diagram in Fig. 1. Our work is just beginning and on the basis of our present results I can only hope to persuade you of the feasibility of studying these processes. In the preliminary studies that we have carried out to date we have as frequently found out about our techniques of experimentation as we have about the processes we have tried to study. Nevertheless, a few findings of more than passing interest have become sharper in the course of our early studies.

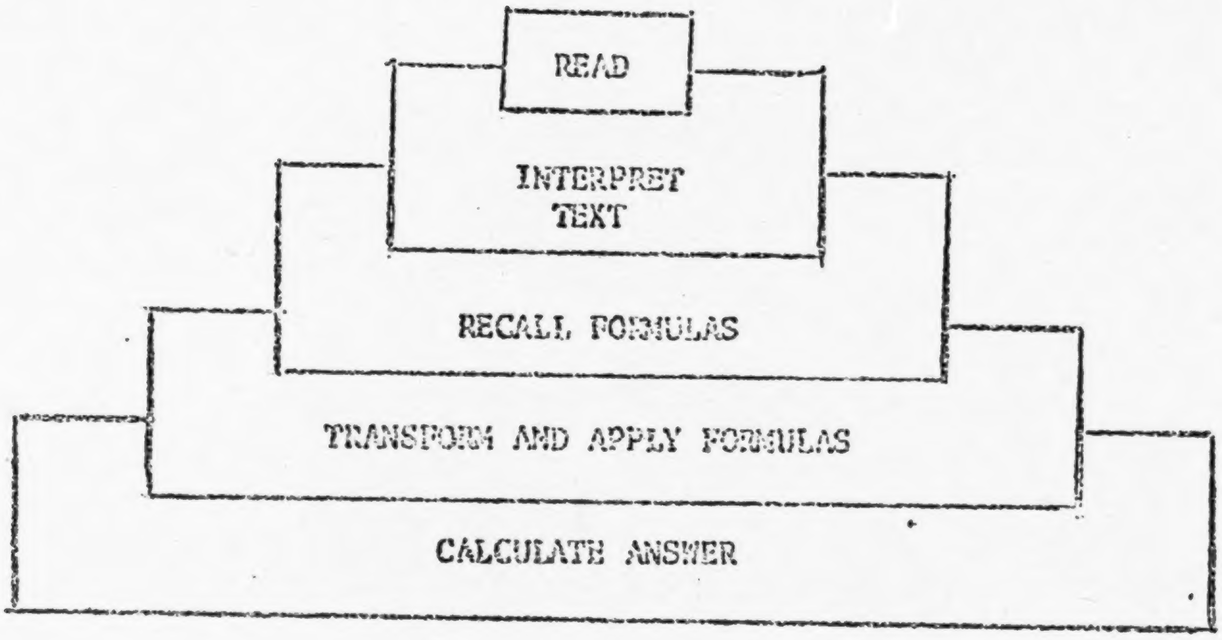

READ
INTERPRET TEXT
RECALL FORMULAS
TRANSFORM AND APPLY FORMULAS
CALCULATE ANSWER

Fig 1

I am particularly hopeful about the prospects for research on these problems because of results obtained recently in the study of processes on either side of those that we have been attempting to study. Work by Clark (1969), Collins and Quillian (1969, 1970), Huttenlocher (1968, 1970), Landauer and Freedman (1968), Meyer (1969), and Trabasso and Rollins (1970) all have contributed new and significant insights regarding the processes of receiving and interpreting information that is related to cognitive structures of a number of kinds. On the other end of the system, work by Dansereau (1969), Groen (1967), Restle (1969), and by Suppes, Hyman, and Jerman (1967) have advanced our understanding of the processes involved in algorithmic calculation. The problem that we have tried to begin to solve is that of understanding the processes that mediate between the interpreted input where concepts involved in a problem and their interrelationships are understood and the process of using an algorithm to calculate the answer. We have supposed that these intervening processes involve memory access in a critical way and that mechanisms for transforming structures that are retrieved from memory into forms appropriate for the problem at hand are also of central importance.

A Study of Response Time for Retrieval and Transformation

The time to obtain a solution for a problem includes the time needed for all of the processes given in Figure 1. It has seemed to me that it would be helpful in the study of memory access and transformations if we could investigate a task that did not require calculation and that minimized the time needed for reading and interpretation.

In an experiment developed and carried out by David Kieras we attempted to meet these criterion. Subjects were given the task of deciding whether or not the value of a variable could be calculated if they were given values of a specified list of variables. Let me use an example; if I were to ask whether the density of a cube could be calculated from its mass and volume the answer would by "yes". Similarly, if the mass and the length of an edge of a cube were given, one could calculate its density. However, if one were given the mass of a cube and the amount of force that was applied to it in some situation, he could not calculate the density of the cube.

In our experiment, subjects began by memorizing a set of four formulas. For one group of ten subjects the formulas described various things involved in a hypothetical automobile highway trip. The formulas were:

$$\text{Driving time} = \text{arrival time} - \text{leaving time.}$$

$$\text{Distance} = \text{driving time} \times \text{average speed.}$$

$$\text{Gas mileage} = \frac{\text{distance}}{\text{gas used}}.$$

$$\text{Total time} = \text{driving time} + \text{preparation time.}$$

The second group of ten subjects also memorized four formulas, but in this group variables were named by letter only. The formulas learned by group 2 were:

$$V = F - L$$

$$D = V \times A$$

$$M = \frac{D}{G}$$

$$T = V + P.$$

Subjects in the experiment had been tested for their ability to solve simple algebraic equations. All of the subjects had a perfect score on a test with six simple problems involving linear equations.

Subjects were given 15 min during which they could study the four formulas in any way they wanted to. At the end of the 15-min period, the experimenter asked the subject to list the permutations of the numbers 1-4. The context provided for this task was that of making up license plate numbers where the numbers 1-4 are the only ones available and the subject could use each digit only once per license plate. This took a few minutes and provided a distracter task.

After the subject completed the permutation task he was asked to recall the formulas. If he was unable to do this, he was asked to review the formulas. All subjects were able to remember the four formulas at the beginning of the main task.

The main task consisted of presenting tabulating cards with a list of variables punched into each card and interpreted for the subject to read. Each card showed the name of a variable, then a dash, and then a list of variable names of length 2, 3, or 4. The typical example for the group with meaningful variable names was:

Arrival time -- gas used, gas mileage, average speed, leaving time.

The correct answer for the card given above was "yes". An item for which the correct answer was "no" was:

Arrival time -- distance, preparation time, gas used.

The items for the group with letters as variables corresponding to the items listed above were:

F -- G, M, A, L,

P -- D, P, G

With nine variables related as they are in these formulas there are 36 problems having positive answers. Twelve positive cases have variable lists of length two, 16 positive cases have variable lists of length three, and eight cases have lists of length four. These numbers exclude problems that have extra variables in the list. For example, the problem

Driving time -- arrival time, leaving time, distance

was not used.

All 36 of the possible positive items were presented along with 36 negative items. The negative items were chosen randomly with probabilities that would match the frequencies of the three list lengths on the average. A block of 72 items was constructed and duplicated giving a total set of 144 items that was shown to a subject. The fact that items in block 2 appeared in the same order and were identical to items in block 1 was noticed by no subjects.

The experimenter conducted the study with a deck of tabulating cards in front of him. On each trial the experimenter turned over a card and started an electric timer. As soon as the subject said either "yes" or "no" the experimenter stopped the timer and wrote the subject's response on the card.

Fig 2 shows the main results of the experiment. The solid lines show mean response times for positive items; the dashed lines are for negative items. The solid points are data from the subjects who used meaningful variable names; the open circles are for subjects with letters used to name the variables. Only correct responses were included in the calculations.

The main result of this study from our point of view was that the experimental situation apparently works. The variables that should produce increased response time apparently did. The effects of list length

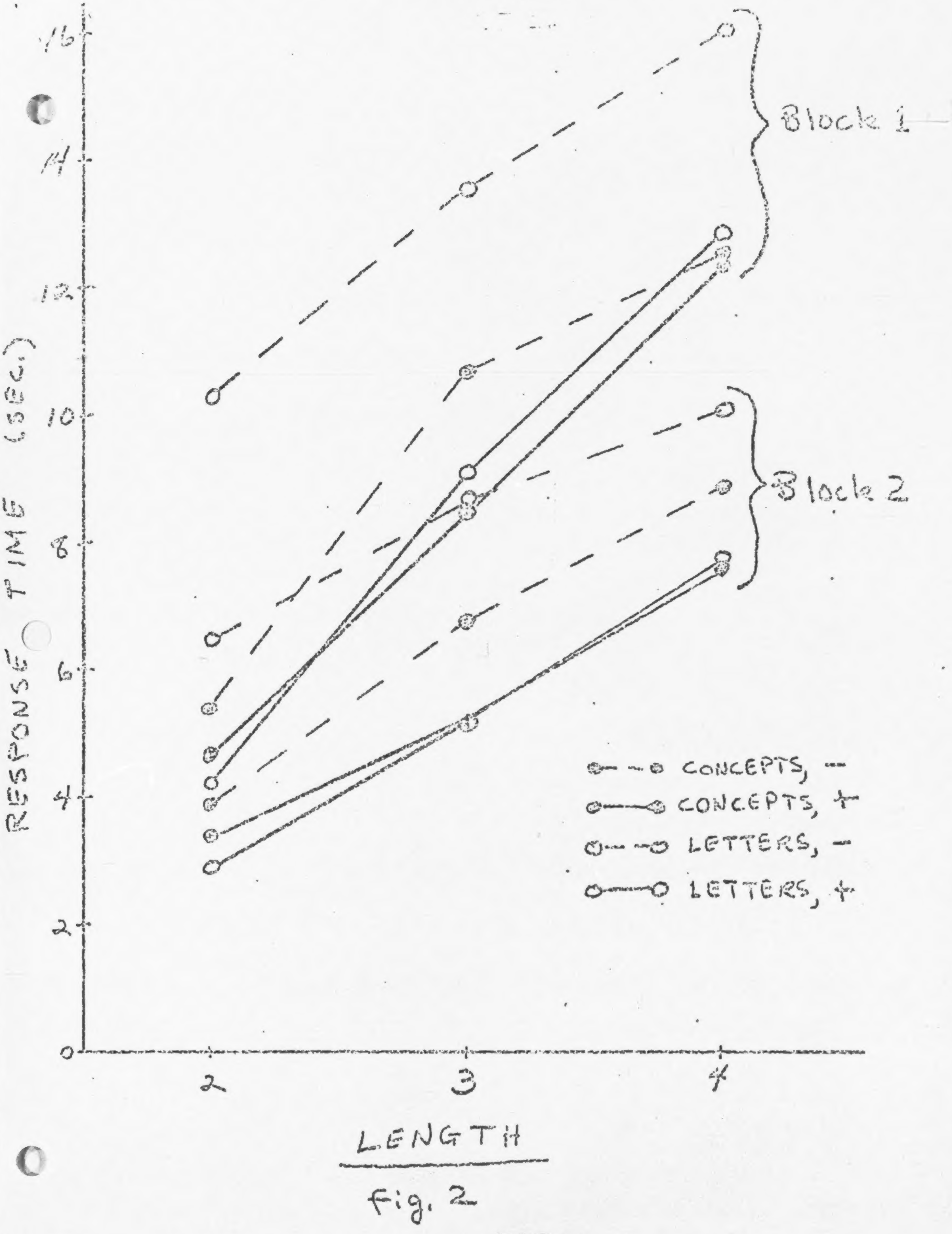

16
14
12
10
8
6
4
2
0
RESPONSE TIME (SEC.)
2
3
4
LENGTH
Block 1
Block 2
CONCEPTS, −
CONCEPTS, +
LETTERS, −
LETTERS, +

Fig. 2

and the difference between conceptual and nonconceptual items were both large and reliable. Furthermore, the technique produced response times that are sufficiently large to permit the inference that most of the response time is time spent in thinking and remembering with only a small fraction of the time spent in reading the variable list and performing the motor response.

Fig 3 shows the proportions of errors made by all the subjects in the second block of trials. Apparently error rates will be something of a problem with this technique. While most of the proportions of wrong response were below .10, nearly one-fourth of the responses were wrong in the hardest condition. We will be able to eliminate one or two subjects from each condition who made many of the error responses, but some experimental refinement will probably also be required. The error rates we obtained can be compared to an average rate of about .08 reported by Collins and Quillian (1969) and results obtained by Meyer (1970) in which error rates were about half as large as those we obtained. In these studies the subject's task was to decide whether a sentence involving two class names was true or false.

The result of greatest interest regarding reaction times was a rather strong interaction between the type of material and whether the answer was positive or negative. Notice that in both blocks of trials, positive answers were given with about the same speed by both groups of subjects. However, negative answers were given much more slowly by subjects who were using variable names that were letters than by subjects who had conceptually meaningful variable names. Further experimental work will be needed to clarify this difference but even this preliminary result suggests an interesting possibility with regard to the process of retrieving structured information from memory.

FILMED FROM BEST AVAILABLE COPY

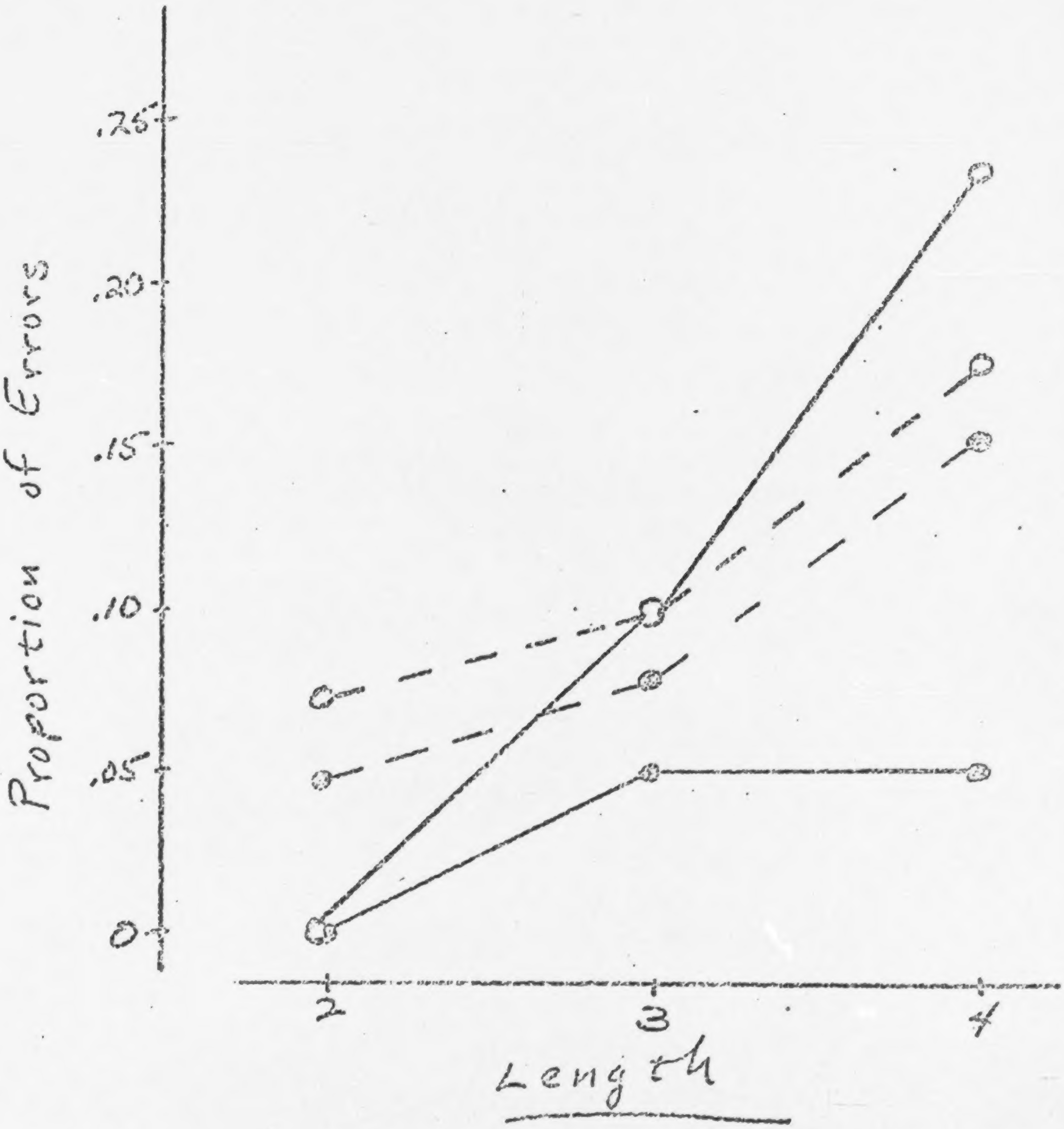


fig. 3

Imagine two subjects working on the task at hand. One of the subjects has a list of formulas in his memory; the other subject has a conceptual network.

The kind of memory structure that the subject with lists might have is given in Fig 4. I suggest that with a list structure a subject would have to perform the present task with some kind of sequential decision process. Fig 5 shows one possible tree for deciding that the problem given below the tree has a positive answer. The lines in the decision tree refer to retrieval of a formula from memory. F occurs in only one formula, so it connects to the two members V and L. L is one of the variables given in the list, so it is circled. The other member retrieved in the first access is V which connects to three pairs of elements. The pair F,L has already been retrieved. The other two might be retrieved from memory either in a sequential or parallel fashion, but the decision to say "yes" could be made as soon as the formula involving V,P and T is recovered from memory.

Fig 6 shows a problem for which the answer is negative and a possible decision tree for arriving at that answer. F connects to V and L as in the problem of Fig 5. L is again in the list. V connects to three pairs of variables as before. This time A is the only one that is in the list that did not occur in the first memory access. This leaves D which connects to two pairs of variables and in one of them there is an element G that is given in the list. However, in order to use G, M would be required. M is not given and the decision is negative. One feature of deciding to say "no" on the basis of a tree like Fig 6 is that all the possibilities have to be exhausted. So apart from the fact that the decision trees are generally more complex for negative problems than for positive ones the subject also has to use more time because he has to exhaust all the possibilities.

$\langle V, F - L \rangle$

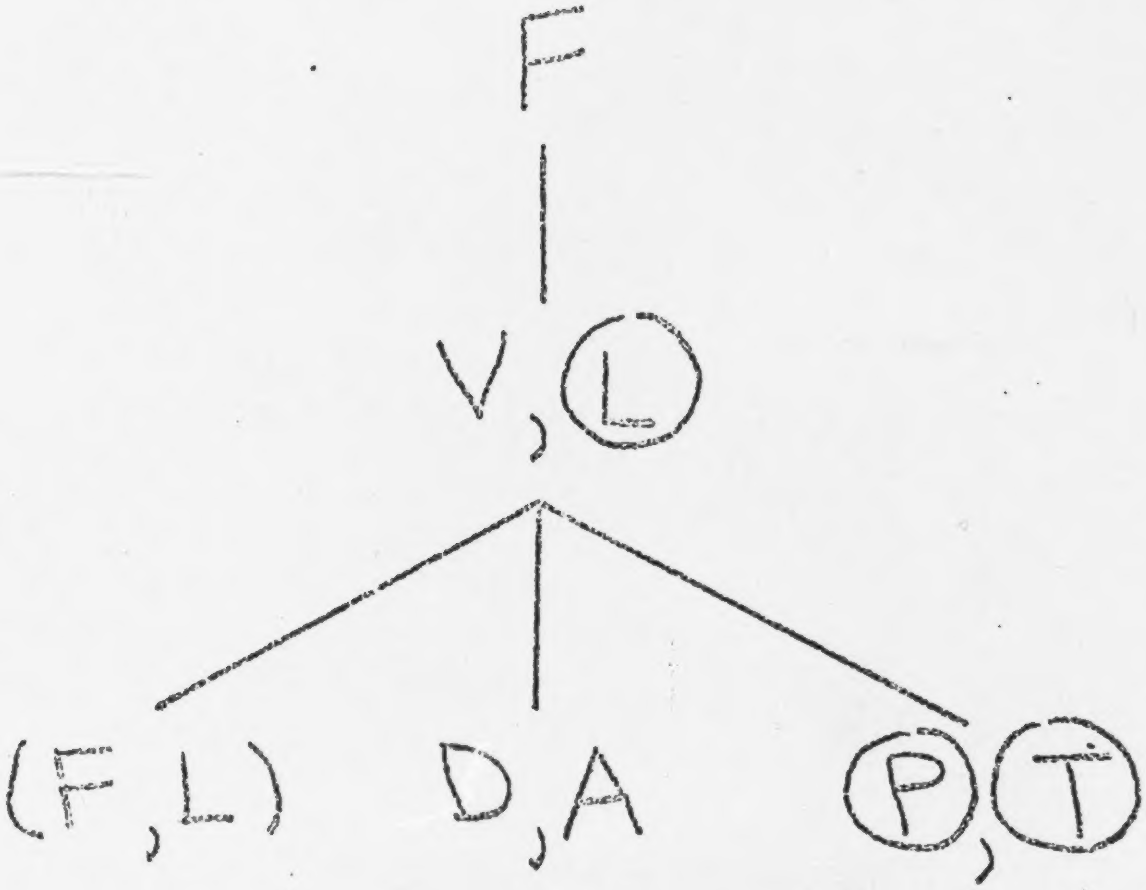
$\langle D, V \cdot A \rangle$

$\langle M, D / G \rangle$

$\langle T, V + P \rangle$

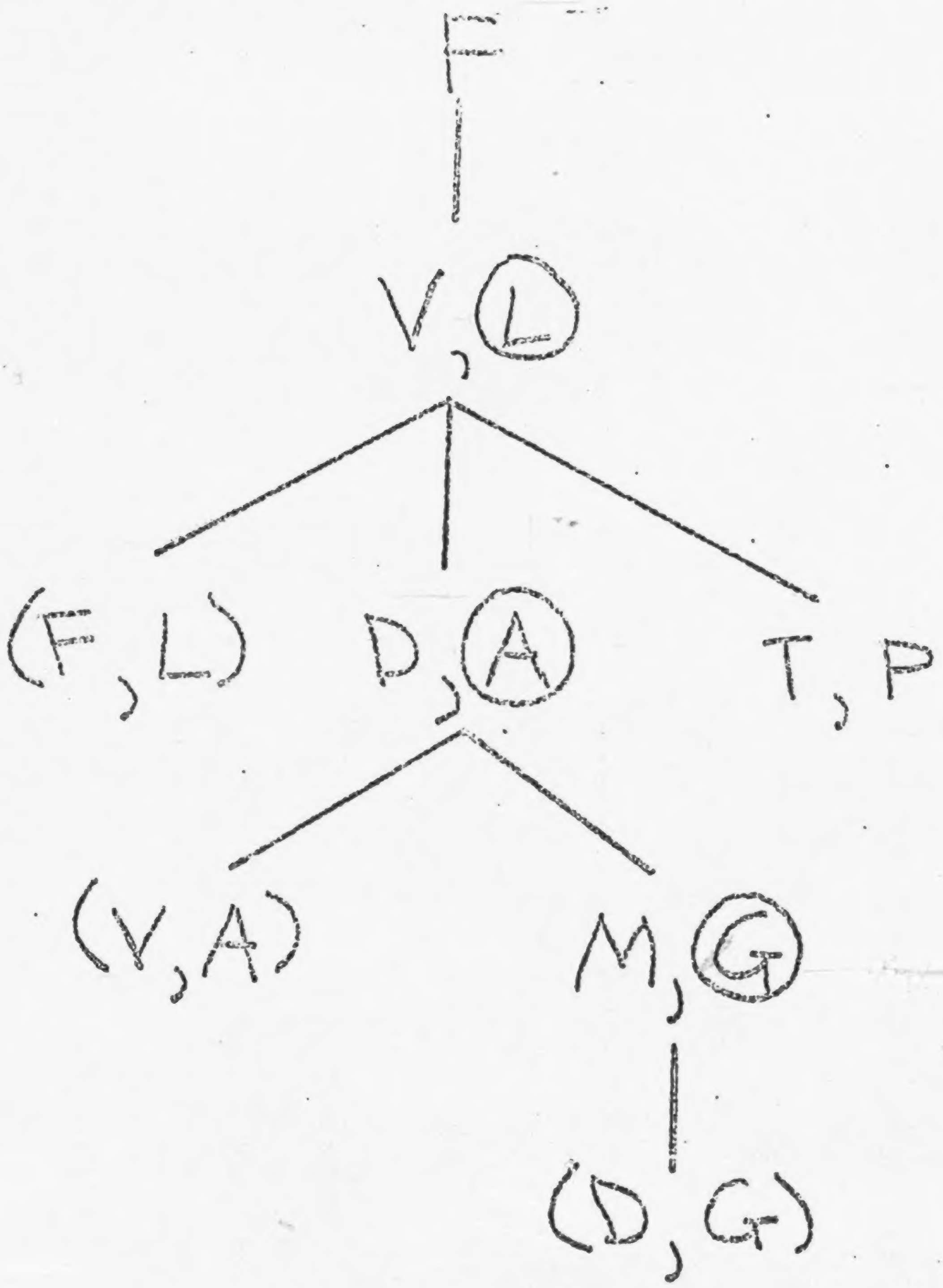
List Structure

Fig. 4



F - P, T, L (+)

fig. 5



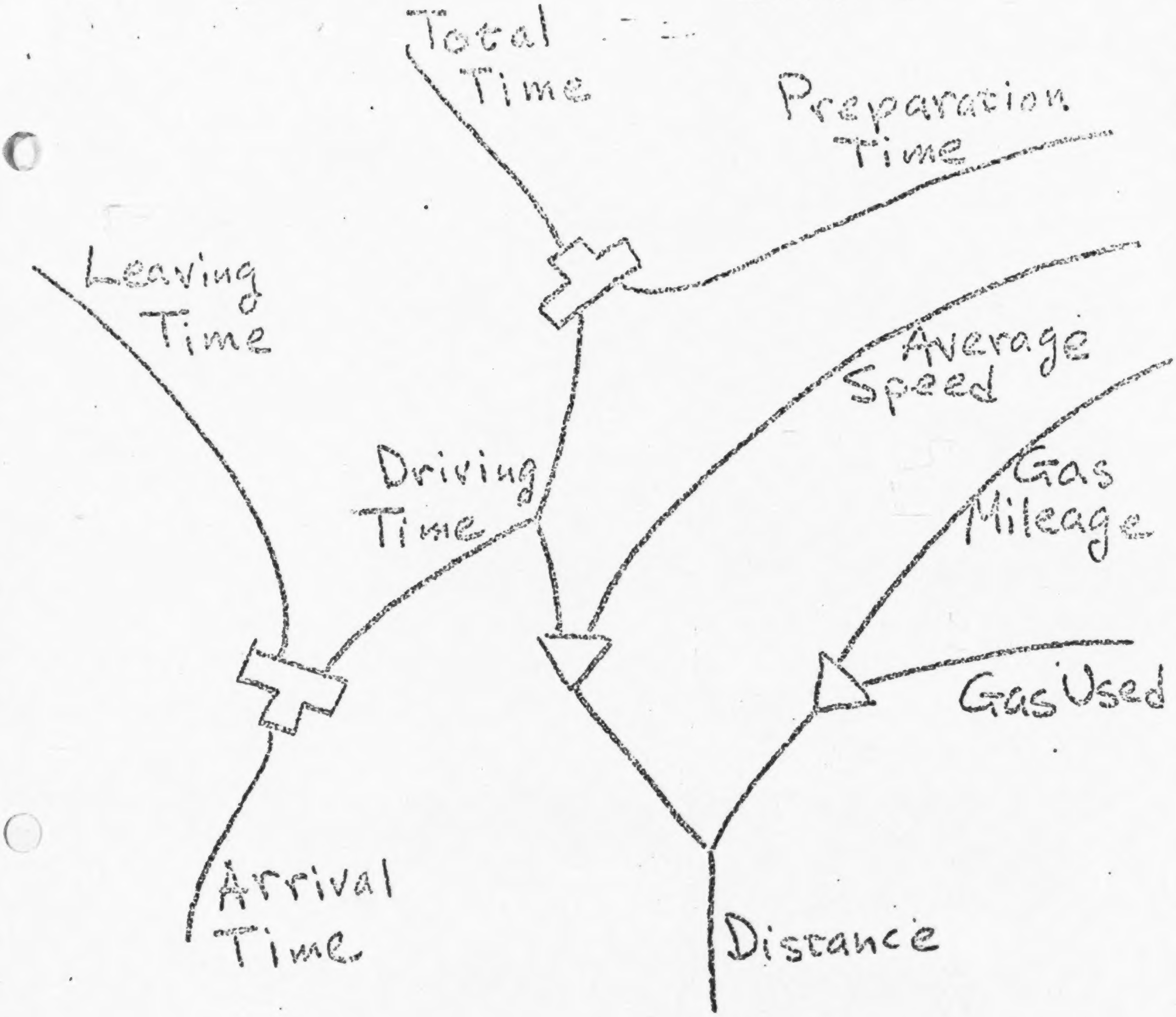
F - L, A, G (-)

Fig. 6

Fig 7 shows a somewhat different structure that might characterize subjects whose problems involved meaningful variable names. This conceptual network has properties in common with Quillian's (1968) analyses of semantic information storage, but the concepts and relations are arranged with the concepts represented by arcs and the relations represented by nodes. This kind of structure has been called a relational network by Reich (1970) who has used such a structural pattern for describing linguistic structures. The difference between the two kinds of operators or relations shown in Fig 7 is simply to represent the operations of multiplication and addition with different symbols. The difference is of no importance in the present content because subjects merely decided whether a set of concepts were connected in any was at all.

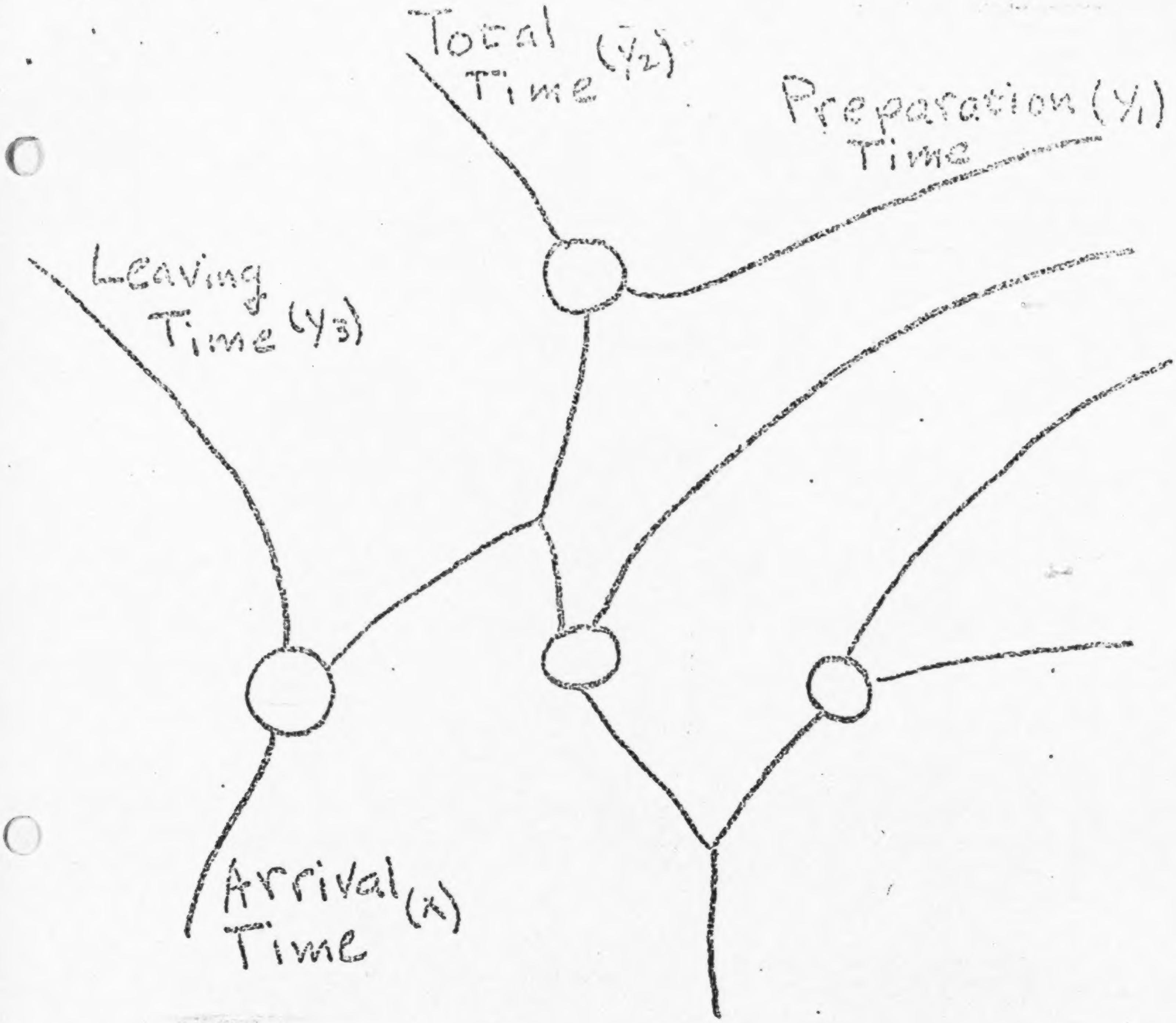
To decide whether a set of concepts is connected, each of the nodes of the graph has to be interpreted as an and-gate. Two elements leading to a node can activate an element reading from it. In Fig 8, I have drawn a schematic graph labelling those arcs that correspond to the concepts in Fig 7 used in the same positive problem as was described by Fig 5. The process of arriving at a correct positive answer would be that of finding a path from y_1 and y_2 into the node above x and noting that y_3 would then activate the final node. Fig 9 shows the corresponding schematic for the negative problem described by Fig 6. The exact process of deciding that the three y arcs do not connect to x is not clear but I, at least, have the intuition that the difference in time to make an appropriate decision given structures like Figs 8 and 9 would be much smaller than the difference in time to make decisions using the list structures described earlier.

Of course, many details of this story have to be filled in and a good



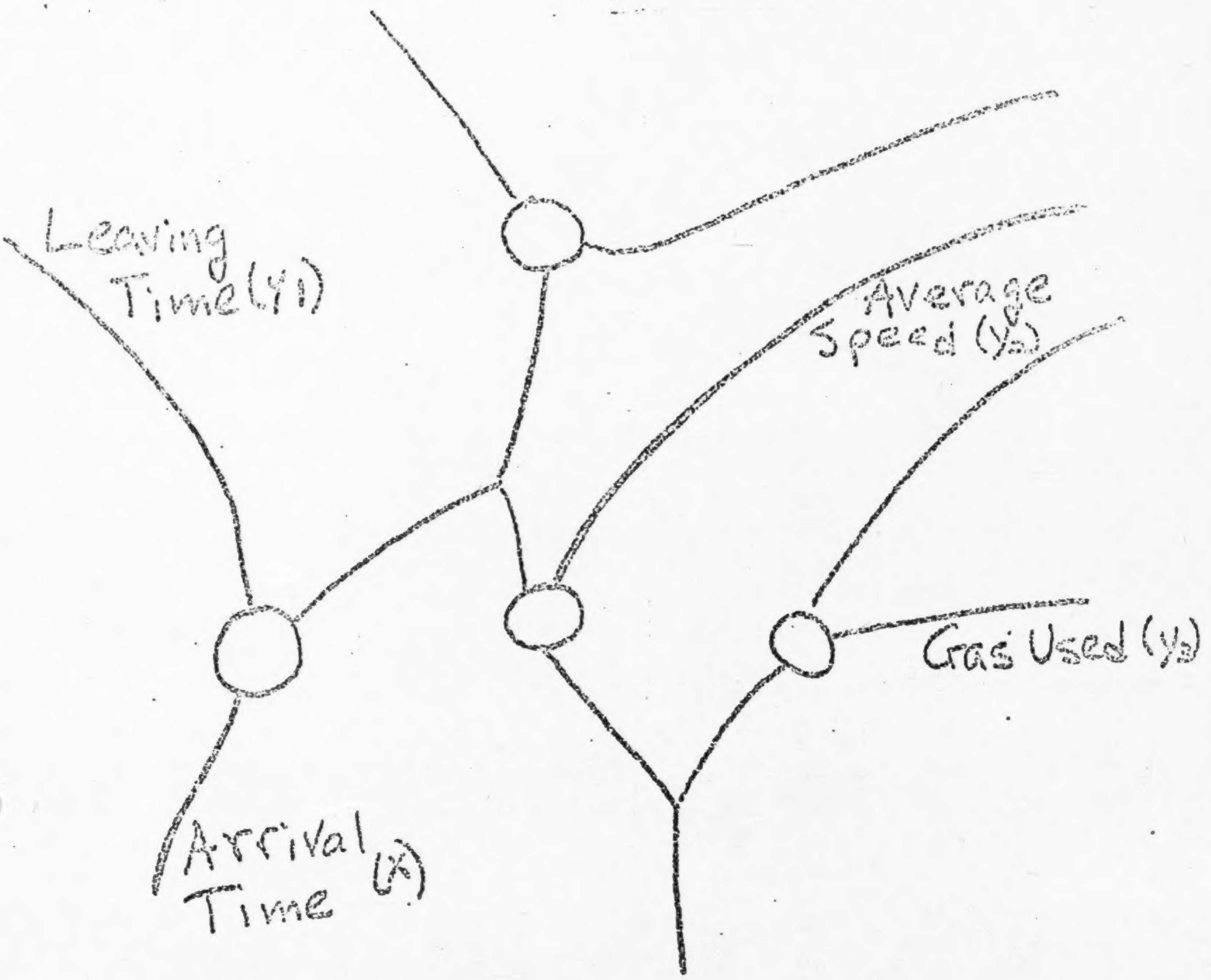
Network Structure

fig. 7



Arrival Time — Preparation Time,
Total Time,
Leaving Time

Fig. 8



Arrival Time — Leaving Time,
Average Speed,
Gas Used

fig. 9

deal of experimental work involving numerous refinements of technique as well as design are ahead of us. Nonetheless I believe that the results that I have presented to you here are sufficient to sketch out the nature of the problems that are involved and to provide some degree of expectation that further work along these lines may well prove fruitful in the development of theory about processes of memory access and transformations performed on structures that are retrieved from memory in problem solving situations.

A Study of the Achievement of Insight

We have been conducting a number of experiments involving the acquisition of cognitive structures. Our experimental material in these studies has been the binomial formula. It has a number of useful properties for our purposes, many of which involve the ease with which our subjects can learn the material, and the convenient fact that practically none of them have learned the material before they arrive in our laboratory. While most of our studies of this situation deal primarily with the learning of a structure rather than with utilization of it, one current study involves the retention and utilization of structures in an interesting way.

This study is being conducted by Douglas Stokes. He began his experiment hoping to investigate the influence of cognitive proximity. This is a notion that Stokes developed to refer to certain effects of context and temporal contiguity in reasoning and thinking. Stokes' idea is that if two propositions are in relatively close proximity in a person's cognitive structure, then it is more likely that he will put them together in some problem situation than if they are widely separated. For example, if some problem or question arises that can be solved or answered only if two historical facts are considered simultaneously, then Stokes believes it is more likely that a person will succeed if both facts were learned in a history class than if the same two facts were learned in different contexts such as a history class and a political rally.

To apply this idea in relation to our studies of simple mathematical structures, Stokes taught two parts of the binomial formula. The two parts are the combinatorial coefficient and the expression specifying the possibility of some sequence. In his original design, Stokes used two different contexts. One of them involved a fictional substance, quixnon, that emits alpha and beta particles and only emits one particle at a time. One instructional sequence contained the formula for calculating the number of sequences with a particular number of alpha particles, and in a separate sequence contained the formula for the probability of any particular sequence of alpha and beta particles. In two different sequences, the same principles were taught using the context of a political candidate running for president who had a probability of winning any one of several states.

The experimental design involved comparison of conditions in which subjects learned about both combinatorial coefficients and probability expressions in the same context with conditions in which subjects learned about principle in the political context and the other principle in the particle physics context. The experimental test involved a problem in which the subject had to calculate the probability of achieving some number of successes in some number of trials. One such problem was, "If a quantity of quixnon emits eight particles what is the probability that four of them will be alpha particles and four of them will be beta particles?" That is, Stokes asked subjects to solve a problem in which they had to use both of the principles or subformulas that they had learned.

Stokes' early result was deflating since no subjects showed any evidence of putting the two ideas together. This was quite surprising to us because most of Stokes' subjects did show that they had learned both substructures that needed only to be recalled and put together in order to solve this problem.

We may return to the question of cognitive proximity in some future study. However, our current experiment represents an attempt to understand the almost total failure of subjects to put together two ideas that appear to be easily related. We think we have the answer, although our results are not in final form as yet. Compare the question stated above with the following: "If a quantity of quirnon emits eight particles, what is the probability that any one sequence out of the set of all sequences which contain four alpha particles and four beta particles will occur?" Our current data include no subjects who failed to answer that question correctly if they successfully completed the preliminary training in which they learned about combinatorial coefficients and probability expressions. (About one-half of our subjects are able to complete our preliminary training successfully.) However, we have yet to find any subject who can solve the problem in the form that does not mention sets of sequences specifically. In fact, some of the performances in this condition are quite bizarre.

The study of insight has always involved the idea that two ideas or structures had to be put together more or less spontaneously by a subject. Kohler's (1927) study of insight in the apes is the best known example. What I think we may have found in our study of insight about the binomial formula is a case in which the bridge needed to connect two ideas is a rather specific one. Apparently, for someone who has just learned the components of the binomial formula, an important condition for seeing the relationship is some mention of a set of sequences. Lacking this, the two components seem to bear no relationship to each other at all However, with an appropriate link subjects seem to perform almost perfectly. One imagines then that a general property of insightful behavior might be the need for some sort of linking or bridging between the component ideas that are brought together in the process of solving a problem.

Patterns of Deductive Reasoning

Studies of incorrect deductive reasoning have been carried out by numerous investigators over a number of years. Many of these studies have emphasized the role of attitudes and feelings in reasoning processes showing that the conclusions that people draw are not entirely free of influence by their attitudes.

A somewhat more cognitive approach has been taken in another study by Douglas Stokes. Stokes chose to study deductive reasoning in a rather naturalistic--almost ethological--experiment. He gave a sheet of paper with several test questions on it to each of a number of subjects. Each item consisted of one statement to be considered as a premise and another statement that will be called a probe. The subject's task was to decide whether the probe followed from the premise and he had three response alternatives. If he thought the probe followed he was to mark "T". If he thought the negation of the probe followed from the premise he was to mark "F." And if he thought that neither the probe nor its negation could be proved from the premise he was to mark "0".

As you would undoubtedly expect, Stokes found that in many cases subjects did not follow the rules of formal logic. But except for a few cases, Stokes' results do not seem to relate to affective or attitudinal processes. Many of Stokes' items involved properties of numbers and many others used nonsense materials. For example, one premise was, "If x is equal to either 1 or -1, then y is equal to zero. If y does not equal 0, then z is a positive number." The probe was, "If x is equal to -1, then z is a positive number." Another item was, "if x quins, then y sprums. x does not quin." The probe for this one was, "y does not sprum." Stokes included other items that probably have some emotional content such as, "If the dog dies Susan will be sad," and "Russia is reactionary and Red China is progressive." But the responses to those items did not seem to be distinguishable from those to the items that had little or no affective content.

The rules of logic can be thought of as a generative grammar in the sense that they can be used to generate all of the valid arguments, and only those arguments that are valid, within a corpus of propositions. The problem that Stokes took on can be thought of as the problem of finding a grammar that will generate the arguments judged to be valid by the subjects in his experiment. Stokes chose the examples in his test in an attempt to get as much information as he could about the deviations between his subjects' grammar and formal logic. He stated his conclusions in relation to the operators of formal logic. I will not try to present all of Stokes' results in detail, but I will try to state and illustrate a few of his main findings.

As many of us probably would expect, a fair number of the deviations between subjects' reasoning and formal logic have to do with the operation of implication. Stokes found three characteristics of subjects' reasoning involving "if,then" statements that depart from conventional logic.

One of Sickes' findings was the one that many students are aware of when they begin learning the correct use of the "if, then" connective. In formal logic, the statement "if p, then q" is implied either by the negation of p or by the affirmation of "q". For example, from the premise, "Telephones exist. The Black Panthers exist" the conditional statement "If telephones exist, then the Black Panthers exist," can be deduced. Or from, "The glum wurns and walds," we can deduce, "If the glum wuras, then it walds". Only six of 20 subjects would assert that the first conclusion was valid, and eight of 20 asserted that the second one was. Of possible interest is the fact that 13 of 20 subjects said that "If the man goes into town, then he will die," follows from, "The man will die". It may be that subjects are less willing to conclude, "If p, then q" from "p and q" than they are from "q" alone.

In any case, subjects seemed to be completely unwilling to derive, "If p, then q," from "Not p". From "Unicorns do not exist," no subjects would assert that "If unicorns exist, then there will never be a nuclear war" could be deduced. And the same finding was obtained with sentences of this form dealing with the bravery of generals, the composition of the moon, and the value of x.

When a conditional sentence follows from its consequent or the negation of its antecedent, subjects tend to count a deduction as invalid when in fact it can be performed. There are other cases in which subjects are willing to accept deductions that are not valid in formal logic. One category of these can be described as a tendency for subjects to map conditional statements into biconditional statements. This has two forms. First, on the basis of the statement "If p, then q," subjects often are willing to conclude that the statement "If q, then p" is a valid deduction.

For example, from, "If Harry smokes, then he is a carbon copy of his father," 14 of 20 subjects agreed that "If Harry is a carbon copy of his father, then he smokes" followed. It is interesting that sentences of exactly the same form can produce quite different results. Stokes found that with the item, "If the temperature is below zero, then the car will not start" only three subjects would agree that "If the car will not start, then the temperature is below zero" followed. But the latter example is an exception to the main gist of Stokes' findings. With most of the sentences that were used, subjects were willing to conclude that "If q, then p" followed from "If p, then q" in a majority of cases.

The other result supporting the idea that subjects map conditional statements into biconditional statements is that from a premise "If p, then q" subjects are willing to conclude "If not p, then not q". For example, from the premise, "If the man is kind then he will give the child a nickel," 12 of 20 subjects asserted that, "If the man is not kind, then he will not give the child a nickel," followed.

The third result of Stokes' experiment involving implication is somewhat more complicated. If the antecedent of a conditional statement is a disjunction ("p or q") and the consequent is some statement ("r") then the whole statement is consistent with, "If q, then not r". For example, "If either the dog or the cat is sick, then we will call the vet" does not contradict, "If the dog is sick, we will not call the vet". However, Stokes' subjects, like many of us, I think, did not understand the sentences that way. 20 of 20 subjects said that the conclusion given above was false on its premise. Stokes concluded that if two statements share a disjunct in their antecedents but have inconsistent consequents, subjects treat the statements as inconsistent.

A fourth property of the reasoning that subjects performed in Stokes' experiment involves the operator of disjunction in conventional logic. The statement "p" implies the statement "p or q", but subjects seemed quite uncertain about making deductions of that form. In one item the premise was "Russia is reactionary and Red China is progressive" and subjects were asked about "Either Russia is reactionary and Red China is progressive or Russia is not reactionary and Red China is not progressive." Only ten subjects said the probe followed; eight said the probe was false. From, "The child's IQ is 150" only six subjects said that "The child's IQ is either 150 or 160" followed, and 11 subjects said the probe was false on the premise.

Another finding, however, was that when the deduction of a disjunction from one of its components is needed as a substep of a larger deduction subjects apparently are willing to assert that the disjunction follows. From the premise "If x equals either 5 or 9, then y is less than 3, If y is less than 3, then z is greater than 0," the probe "If x is equal to 5 then z is not greater than 0" was judged false by 19 of 20 subjects. From the premise, "If the native is from the hills or the plains, he can tell us where the stone is hidden. If he can tell us where the stone is hidden then our mission will be a success," 18 of 20 subjects were willing to assert that the probe, "If the native is from the hills, then our mission will be a success" followed.

The general findings of Stokes' study consist of a set of rules that he concludes are characteristic of subjects' deductive reasoning. Some examples are:

"If p, then q" tends to be interpreted as "p if and only if q."

"If p, then q" often does not follow from "q" and never follows from "Not p."

the case of the grammar of deductive thinking we are in a situation not unlike the one faced by investigators who attempt to discover the grammar of a general language. To be sure, there is a well defined grammar of deductive thinking, that is, formal logic. But it is not surprising that subjects do not follow that grammar in all cases. One of the reasons for this is probably that while internal consistency is a compelling criterion for logicians consistency is less compelling in ordinary situations than the criterion of usefulness in interpreting events that occur. The study of structures that are poorly defined in this sense is of extreme importance by such studies face certain problems that are avoided when cognitive structures correspond closely to well-defined formal structures.

Conclusions

I have tried to present a picture of our first tentative steps in a program of research on a relatively new problem. All three of these studies has dealt with some aspect of the process by which a subject uses relationships between elements of his cognitive structure. In Kieras' study, we investigated ways in which subjects use relationships between mathematical formulas that share certain variables. Stokes' experiment on deductive reasoning dealt with relationships among sentences that share concepts. And Stokes' experiment on insight is concerned with substructures that can be combined into a single, more complex formula.

The results that we have obtained in these early studies do not fit together in any tidy, coherent way. We will have to wait for another occasion to consider a general theory about retrieval and transformation of cognitive structures in problem solving. On the other hand, the results that we have obtained permit me to hope that the waiting time until such a theory can be proposed will not be infinite.

References

Clark, H. H. Linguistic processes in deductive reasoning. *Psychological Review*, 1969, *76*, 387-404.

Collins, A. M. and Quillian, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1969, *8*, 240-247.

Collins, A. M. and Quillian, M. R. Experiments on semantic memory and language comprehension. In Gregg, L. W. (Ed.), *Cognition in learning and memory*. New York: Wiley, 1970.

Dansereau, D. F. An information processing model of mental multiplication. Doctoral dissertation, Carnegie-Mellon University, 1969.

Groen, G. J. An investigation of some counting algorithms for simple addition problems. Report No. 118, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.

Huttenlocher, J. Constructing spatial images: a strategy in reasoning. *Psychological Review*, 1968, *75*, 550-560.

Huttenlocher, J., Higgins, E. T., Milligan, C., and Kauffman, B. The mystery of the "negative equative" construction. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 334-341.

Kohler, W. *The mentality of apes*. New York: Harcourt, Brace, 1927.

Landauer, T. K. and Freedman, J. L. Information retrieval from long-term memory: category size and recognition time. *Journal of Verbal Learning and Verbal Behavior*. 1968, *7*, 291-295.

Meyer, D. E. On the representation and retrieval of stored semantic information. Report MMPP 70-3, Michigan Mathematical Psychology Program, The University of Michigan, 1970

Quillian, M. R. Semantic memory. In Minsky, M. (Ed.) *Semantic information processing*. Cambridge, Mass.: MIT Press, 1968. Pp. 216-270.

[illegible], P. Doctoral dissertation in preparation. The University of Michigan, 1970.

Restle, F. Speed of adding and comparing numbers. Report 69-5. Indiana Mathematical Psychology Program, Indiana University, 1969.

Suppes, P., Hyman, L., and Jerman, M. Linear structural models for responses and latency performance in arithmetic. In Hill, J. P. (Eds.), _Minnesota Symposia on Child Psychology_. Minneapolis: Minnesota, 1967. Pp. 160-200.

Trabasso, T. and Rollins, H. Storage and verification stages in processing concepts. Mimeographed manuscript, 1970.

Semantic Structure and Clustering in Free Recall[1]

by

W. Kintsch, R. F. Miller, and R. M. Hogan
University of Colorado

Abstract

If output order in the free recall of word lists is in part determined by the semantic relationships that exist among the words of the list, the cluster structure that is obtained in free recall should be closely related to the cluster structure that is obtained in other tasks which depend strongly upon the semantic organization of long-term memory. This hypothesis was tested in an experiment comparing the organization of word lists in free recall with the organization of the same lists in a sorting task, a word identification task, and an association task. Cluster analyses of the data revealed that whatever organization was present in the various tasks tended to be highly similar.

In this paper we investigate the extent to which the general structure of memory determines the output order in the free recall of word lists. There are a number of factors which are known to affect recall order. First, there is the input order which is perhaps the strongest determinant of output order. In fact, in multi-trial free recall experiments in which a constant input order is used the output order conforms to the input order after some trials (Mandler and Dean, 1969). A second factor that is known to determine the output position of an item on any given trial is the output position of that item on previous trials. More specifically, items that have not been recalled previously tend to be recalled early; as an item is recalled more often it moves back in the output order (Battig, Allen and Jensen, 1965; Mandler and Griffith, 1969; Roberts, 1969). Finally, items cluster in free recall because some meaningful relationship exists between them: words that are strongly interassociated, words that belong to the same taxonomic category, as well as adjective-verb-noun triplets that are meaningfully related all tend to be recalled next to each other (see Kintsch, 1970a, for a review of association and categorical clustering, and Stanners, 1969, for grammatical clustering). The problem that will be considered here is what mechanism produces this kind of clustering. The important factor that seems to be common to all of the examples given above is that the items that cluster are meaningfully related. Indeed, if the meaning relations are destroyed, e.g. by presenting nouns together with inappropriate adjectival modifiers, clustering is disrupted (Cofer, Segal, Stein, and Walker, 1969). The meaning of a word must be represented in memory somehow. We shall call the component of long-term memory that establishes semantic relationships among words the lexical component of memory. Furthermore, we propose the working hypothesis that this lexical component of memory provides the basis for clustering in free recall. Alternatively, clustering in free recall could be caused by factors specific to the free-recall task, rather than the general memory structure. The experiments reported below were designed to differentiate between these two hypotheses.

Of course, attributing clustering to the "lexical component of memory" or to "meaning" is in itself not more illuminating than talking about "associative habits", "grammatical habits", or "organization". It is necessary to specify more precisely what is meant with lexical structure. Two approaches to this problem have been tried. One is to work out a model of memory structure from which, eventually, testable deductions can be derived. Some preliminary results have been reported elsewhere (Kintsch 1970b). In the present paper we are trying to approach the problem differently. Although we do not know how the lexicon is structured, and we do not have a model for it, we can compare the output order in free recall with performance in various other tasks in which lexical structure may reasonably be expected to be a crucial factor, in particular in a sorting task, a word-guessing game, and restricted associations. Although the limitations of this empirical approach are obvious, it can provide evidence as to the general adequacy of our working hypothesis. If the structure of the output order in free recall does not correlate well with that obtained in the other tasks mentioned, the usefulness of the hypothesis that the output order in free recall reflects, inter alia, lexical structure would be questionable.

One way to obtain some information about the lexical structure is to ask Ss to sort a set of words into subjective categories on the basis of the similarity of their meaning. In order to do so Ss must rely on the word dictionary that is part of their long-term memory store. Presumably, similarity judgments reflect the semantic distance between words in that dictionary. We need not be specific here about how this "semantic distance" is to be determined. Miller (1969) has discussed some possible models and their implications for sorting tasks. For present purposes it is merely necessary to assume that category sorting provides a measure of the relatedness between words: words that are sorted into the same category by many Ss are more related than words that are sorted together by few Ss.

A second estimate of how words are organized in memory can be obtained by observing performance in a word-identification task. In a popular parlor game called 20-questions one participant is allowed to ask up to 20 questions in trying to identify a word selected by a second player. The questions asked in this situation reflect the inquirer's memory structure, i.e. the way in which his word dictionary is organized. One can think of Ss' performance in the 20-question game as a search through memory for the "right" word. Starting with general considerations Ss ask questions which permit them to restrict the range of possible alternatives. Clearly, the kinds of questions that are asked depend upon how the information about words is organized in memory. There are of course other factors that influence what questions are asked, e.g. the nature of the search strategies used by Ss. However, it seems reasonable to assume that the sequence of questions asked in identifying a word represents a path through the memory structure to the word in question, although not all way stations on this path may be explicit, and although there may be discontinuities in this path as S abandons one line of inquiry and tries another. One can attempt to estimate memory structure from the nature of the questions asked in the 20-question game. Suppose we have a list of the questions that were asked and answered "yes" in identifying a word a, and a second list of questions that were asked and answered "yes" in identifying a word b. We shall take the number of questions that appear on both lists as a measure of the proximity in memory between a and b. For instance, if a were man and b were woman the questions asked in identifying these words would be mostly identical, and a high proximity measure would be obtained; if a were tree and b conscience, there would be very little or no question overlap and a correspondingly small proximity measure would be obtained. Note that the assumption that question-overlap varies inversely with the semantic distance between to-be-identified words need not hold in every specific case, in that it is not too difficult to devise non-overlapping question sequences for closely related words, or strongly overlapping sequences for quite unrelated words. However, every subject traces his own search path through memory and on the average semantic distance and question-overlap should be related in the way we have postulated.

A third indication of the nature of the semantic relationships among words was obtained from the response overlap in an association task. Suppose R(a) is the set of associative responses that were given by a group of Ss to the stimulus word a, and R(b) is the set of responses to stimulus word b. We shall assume that $n(R(a) \cap R(b))$, i.e. the number of associations that a and b have in common, is a measure of the semantic distance between the words a and b. For this part of the experiment no new data need to be collected, but recourse may be taken to existing association norms. The norms used here were the Michigan Norms of Riegel (1965). These norms present data on 16 restricted association tasks for each stimulus word, thus allowing a more thorough evaluation of similarities among words than would be possible from free association norms.

Thus, three different measures of the strength of semantic relationships among a set of words may be obtained. The three measures are clearly not equivalent, since task demands are quite different in the three cases. However, there ought to be a substantial overlap among them in that they all depend in some way upon how the lexical component of memory is organized. A fourth measure among the same set of words will be based upon the degree of clustering in free recall, and it is the experimental hypothesis that whatever clustering in free recall is obtained will be largely predictable from the organization of Ss' word memories, as inferred from the sorting, word identification, and association tasks described above.

In all experiments to be reported here data were obtained by pooling the performance records of many Ss. Idiosyncrasies of the memory structure are lost by such pooling of data. However, it is reasonable to assume that there exists considerable communality in the memory systems of our Ss, and it is this portion of the memory structure which is common to all or many Ss that can be studied with the present procedures.

EXPERIMENT I

Method

Subjects. All Ss were undergraduates at the University of Colorado who were fulfilling a course requirement. Fifty-one, 34, and 59 Ss served in the free recall task, the sorting task, and the 20-question game, respectively. All Ss participated in individual sessions.

List construction. Forty nouns were selected from the norms of Riegel (1965), five words from eight different categories each. Six concrete noun categories and two abstract noun categories were used from those described in the norms. One word had to be deleted because of a typing error (mutton was mistyped as muffin) in some lists. The actual word list is given in Table 1. Each word was typed in the center of a small index card.

Procedure: Free Recall. Ss went through the stack of cards at their own pace, reading the word on every card and turning the card over. The stack of cards was reshuffled on every trial for every S in order to randomize order effects. After the last card, E gave S a three-place number and S started counting backwards from that number for 15 sec. to the beat of a metronome. After that S was asked to recall as many words as he could without regard to order. The 15 sec. delay between study and recall was introduced in order to avoid recall from a short-term recency buffer, which could be independent of semantic structure. Unlimited recall time was given. Each S received five trials.

Sorting. S was given the cards in random order and a board with eight bins marked on it. The instructions were to sort the cards into from two - eight categories according to their similarity of meaning. Cards were sorted one by one, and Ss were not allowed to make changes or to retrace their sorting. Only the top cards in each category were visible at all times. The task was self-paced. After finishing, S was asked to reorder the deck, making any changes that seemed indicated, now that he knew more about the population of stimuli to be stored.

Twenty Questions. Every S worked on as many problems as he could during a one-hour session. Ss never received more than one word from a conceptual category, but otherwise the words to identify were selected at random from a pool of available words. A word remained available until it had been identified three times, or until it had been tried five times. Ss wrote their questions in a format that emphasized the binary nature of the question: _________ OR _________ . The E then circled that part of the question that was correct, which in some cases may have been both or neither part of the question. The S was allowed to see his old questions and what answers were given to them throughout the time that he was working on a problem. A problem was terminated either upon solution or after 20 questions had been asked. The task was self-paced.

Restricted Associations. The Michigan norms (Riegel 1965) list the responses of 100 Ss to 200 stimulus words on 16 different restricted association tests (superordinates, coordinates, subordinates, similars, contrasts, nouns, verbs, adjectives, foregoing words, following words, locations, wholes, parts, precedings, contemporaneities, succeedings). All responses that occurred with a frequency greater than one were tabulated. These tabulations for the 40 words used here provided the data for the present experiment.

Results

The basic data for each task consist of a proximity matrix that shows for all word pairs the measures of relatedness that were obtained in this task. The 39 words of the experimental list appear in the rows and columns of such a matrix and the entry in the i^{th} row and j^{th} column is the measure of proximity between words $\underline{i}$ and $\underline{j}$.

Free Recall. Free-recall proximity matrices were constructed by counting how often each word was followed by every other word in the recall protocols of 48 Ss. The data from the 1st and 5th recall trial were analyzed in this way. Three Ss were excluded from the analysis. These Ss recalled all 39 words perfectly in the input order after just one self-paced presentation. Obviously, these Ss were not clustering items semantically and hence their data could not contribute anything but noise to the present analysis. Their performance emphasizes the importance of input order as a determinant of output order in recall, but for an analysis of semantic clustering their data are useless. Of course, many other Ss recalled partly in the input order, but this kind of "noise" is unavoidable in an experiment like the present one: one cannot avoid input effects, but merely average out such effects by using many different input orders.

Taking the number of times that one word was followed by another word as a measure of proximity is somewhat arbitrary, in that other reasonable measures could easily be devised. (e.g., one need not limit a proximity measure to items that are immediately adjacent.) However, the measure used here has the advantage of being the most straightforward and simplest one.

Sorting. The number of Ss who put two words together in their sorting was used to generate the sorting-proximity matrices. Separate proximity matrices were obtained for Sort 1 and Sort 2.

Twenty Questions. For each item a list of all questions that were asked and answered "yes" in identifying the item was constructed. All such questions for all words in the list that could be regarded as synonyms or paraphrases were recorded in a "dictionary" and this dictionary was then used to standardize the question lists associated with each item. What constituted a paraphrase was determined by the agreement of the judgments of two Es. The standardized question lists were then compared with each other for all items, and the number of questions common to every pair of items was noted and entered into a question-proximity matrix.

No solutions could be obtained for 6 words; the remaining 33 words were identified either two or three times each. These differences in solution frequencies constrain the proximity values that can be obtained. Items for which no solutions were obtained will not be clustered by the cluster analysis. However, note that the same situation exists, though to a much less serious degree, in the free recall matrices since not all words were recalled equally often. Thus, the estimates of organization that are obtained from the free-recall and especially the 20-question data are biased towards items that are easily identified or easily recalled.

Restricted Associations. For each word pair the number of common responses in the Michigan Norms was computed. Computations were performed separately for each of the 16 association tasks. For instance, the response SKY was given to the stimulus BUTTERFLY as well as to the stimulus BIRD on the Locations task; this was counted as one association overlap. The final association-overlap score for each word pair was the sum of all overlap scores on the 16 restricted association tasks. These scores were entered in the association-proximity matrix. The total number of responses for each word was regarded as the overlap score of each word with itself.[2]

Cluster Analyses. It is not easy to compare quantitatively the structure of the proximity matrices thus obtained. Merely correlating them gives a very poor idea of the amount of similarity present, because the structure or organization that exists in these matrices is determined by relatively few of the many proximity values, with most entries being zero or close to zero. A much more satisfactory procedure is to obtain estimates of the organization of the data from the proximity matrices by means of extant statistical techniques and then compare the resulting structures. The technique employed here is the BC TRY cluster analysis (Tryon and Bailey, 1970). The BC TRY system is a method that, given a matrix of proximities, generates a clustering of the input items according to a standard set of criteria. How many clusters will be generated, how large these clusters are, how they are interrelated, and how many of the items will be clustered depends jointly upon the input matrix and the internal fixed criteria of the BC TRY System. In principle, the analysis proceeds as follows. First, an intercorrelation matrix is obtained from the proximity matrix, i.e. the proximity between items _i_ and _j_ is replaced by intercorrelation between the i^{th} and j^{th} column. This correlation can be regarded as a derived proximity measure, which takes into account not only the proximity between _i_ and _j_ but the whole pattern of relationships of _i_ and _j_ to the remaining variables. Next, one variable is selected as a "pivot variable" and other variables are clustered with this pivot variable if they exhibit a similar pattern of relationships with all other variables as the pivot variable. Similarity is measured by a "coefficient of colinearity" and there are fixed cut-off points below which variables are no longer clustered. Further clusters are formed from the residual correlation matrix according to the same principles. The analysis is fully described in Tryon and Bailey (1970).

The only thing important for the present application of the BC TRY analysis is that it provides an objective method to determine the cluster structure of a set of items, and that it does so without making too many assumptions either about the nature of the data or about the nature of the organization which is to be detected. It simply combines related elements into clusters, where "related" is defined in terms of the pattern of interrelationships among all items. In the present application we are not at all concerned with what these clusters are, whether they make sense, what the geometry of the cluster space is like, or even how much of the variance the clusters account for. We are working with an arbitrarily selected word list, and such information would have no generality at all. We merely want to compare the cluster structure obtained from the sorting data, the word-identification data, the association data, and the memory data. The only question is whether the clusters are the same that are obtained from these procedures; what they are and how good they are is not of much interest.

The results of the cluster analyses are shown in Table 1.[3] The order in which the clusters are listed for each analysis is not significant. A cluster consists of the words for which a check mark appears in the same row. E.g. cluster g in the analysis Sort 1 consists of the two words BABY, GIRL . Whenever a column has no checkmark in any of the rows of an analysis the corresponding item was not clustered at all by that analysis. For instance, SEX was the only single in the Sort 1 analysis. The proportions of items that were clustered in each analysis are shown in the top part of Table 2. These values indicate how strong the organization was in the different sets of data. With the exception of the first Memory analysis these values are quite high. The value shown for the 20-question game was computed for only those items that had been identified by at least two Ss. The six words which were never identified in the experiment and which had to be omitted from the analysis are indicated by zeros in Table 1.

Some similarities and differences among the clusterings that are shown in Table 1 are immediately obvious. For instance, some analyses, especially the 20-question analysis, produce only a few large clusters, while others, especially Memory 5, generate many smaller clusters. However, frequently the smaller clusters are merely a subdivision of a large cluster. For instance, Sort 1 and Sort 2 are identical except that where there are two clusters in Sort 1--BABY, GIRL and DOCTOR, NURSE, PROFESSIONAL - there is only one in Sort 2, which also includes SEX, a word that was not clustered in Sort 1.

In order to compare the clusterings shown in Table 1 with each other a quantitative measure of similarity is needed. For present purposes, it seems desirable to distinguish between differences in the over-all strengths of two clusterings and differences in the cluster structure, independent of over-all strength. Given different task demands one surely could not expect that the effects of organization manifest themselves with equal strength in sorting, problem solving, or memory. Rather, we wish to investigate the hypothesis that whatever organization there is is closely related for the various tasks. In other words, in order to evaluate the data presented in Table 1, two considerations should be kept separate: first, how many items are clustered by each analysis, and second what proportion of those items that are clustered by any two analyses are clustered alike?

Suppose that there are two clusterings of the same set of N elements

$$\underline{\underline{A}}^* = \{A_1, A_2, \ldots\ldots, A_J\} + \{X\}$$

$$\underline{\underline{B}}^* = \{B_1, B_2, \ldots\ldots, B_K\} + \{Y\}$$

where the A_i's and B_i's are clusters of two or more elements and $X = \{x_1, \ldots, x_a\}$, $Y = \{y_1, \ldots, y_b\}$ are sets of single elements. Clusterings are defined such that $A_i \cap A_j = B_i \cap B_j = \emptyset$ if $i \neq j$. Let $n(A_i)$ be the number of items in cluster A_i. Obviously $\sum_{i=1}^{J} n(A_i) + n(X) = N$. The fewer single items there are in a clustering, the stronger the over-all organization; hence $\frac{N - n(X)}{N}$ is an index of the over-all strength of organization for clustering $\underline{\underline{A}}^*$. Now consider the comparison between $\underline{\underline{A}}^*$ and $\underline{\underline{B}}^*$. First of all, we want to know how many items were clustered in common by both analyses, i.e. $N - n(X \cup Y)$. This number is, of course, greatly affected by the individual strengths of organization in $\underline{\underline{A}}^*$ and $\underline{\underline{B}}^*$, but even if this number were very small, whatever clusters $\underline{\underline{A}}^*$ and $\underline{\underline{B}}^*$ may have in common may still be quite similar. In order to obtain a measure of this similarity the reduced clusterings $\underline{A}$ and $\underline{B}$ must be considered where all items that were not clustered by both analyses have been disregarded; in the following equation, and throughout the remainder of this paper, the A_i's which comprise $\underline{A}$ will refer to the reductions of the similarly labeled clusters in $\underline{A}^*$:

$$\underline{A} = \{A_1, A_2, \ldots\ldots, A_J\}$$

$$\underline{B} = \{B_1, B_2, \ldots\ldots, B_k\}$$

Note that $\sum_{i=1}^{J} n(A_i) = \sum_{i=1}^{K} n(B_i) = N - n(X \cup Y)$.

Two clusterings $\underline{A}$ and $\underline{B}$ are identical ($\underline{A} = \underline{B}$) if for every cluster A_i there is a cluster B_j such that $A_i = B_j$, and for every B_j there is an A_i such that $B_j = A_i$. Define the intersection C_{ij} such that $C_{ij} = A_i \cap B_j$ and C_{ij} is non-empty. Two clusterings $\underline{A}$ and $\underline{B}$ are non-identical with respect to C_{ij} if $\underline{A}_i \neq \underline{B}_j$. Consider each set C_{ij} for which A and B are non-identical and determine the minimum number of words that must be deleted in order to establish identity. Let $n(C_{ij})$ be the number of elements in C_{ij}. Let $n(C'_{ij})$ be the number of elements remaining in C_{ij} after deletions have been made. Then

$$S^*_{AB} = \frac{\sum\limits_{\text{all } i,j} n(C'_{ij})}{\sum\limits_{\text{all } i,j} n(C_{ij})} .$$

S^*_{AB} is the proportion of items that are clustered identically by $\underline{A}$ and $\underline{B}$, given that they were clustered by both analyses, and might be taken as a measure of the similarity between $\underline{A}$ and $\underline{B}$. However, this measure is not satisfactory, because clusterings may be quite similar without being identical in the sense that one clustering is a subclassification of the other. For instance in Example (1) $\underline{B}$ is a subclassification of $\underline{A}$, while in Example (2) different clusters are formed by $\underline{A}$ and $\underline{B}$:

(1) $\underline{\underline{A}} = \{abcd,\ efg\}$

$\underline{\underline{B}} = \{ab,\ cd,\ efg\}$

(2) $\underline{\underline{A}} = \{ab,\ cd,\ efg\}$

$\underline{\underline{B}} = \{ac,\ bd,\ efg\}$

However, in each case two items must be deleted in order to establish identity, hence $S^*_{AB} = 5/7$ in both cases.

A more appropriate similarity measure can be constructed if the criterion for deleting items is not identity but set inclusion. If a clustering $\underline{\underline{A}}$ is a subcategorization of $\underline{\underline{B}}$, i.e. if every cluster A_i is included in some cluster B_j, we shall say that $\underline{\underline{B}}$ includes $\underline{\underline{A}}$, or $\underline{\underline{A}} \subseteq \underline{\underline{B}}$. In terms of the inclusion criterion, two clusterings $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are *alike* with respect to C_{ij} if $A_i \subseteq B_j$ or $B_j \subseteq A_i$; $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are *contradictory* with respect to C_{ij} if neither $A_i \subseteq B_j$ nor $B_j \subseteq A_i$. Consider each set C_{ij} for which $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are contradictory and find the minimum number of words that must be deleted in order to remove all contradictions. Let $n(C'_{ij})$ be the number of elements remaining in C_{ij} after deletions have been made. Then

$$S_{AB} = \frac{\sum\limits_{\text{all } i,j} n(C'_{ij})}{\sum\limits_{\text{all } i,j} n(C_{ij})}\ .$$

is a measure of similarity between the clusterings $\underline{\underline{A}}$ and $\underline{\underline{B}}$. S_{AB} is the proportion of items that are clustered alike by $\underline{\underline{A}}$ and $\underline{\underline{B}}$, given that they were clustered by both analyses, where alike is defined on the basis of the inclusion relation rather than the identity relation.

Consider the following example:

$\underline{\underline{A}} = \{ab,\ cd,\ efgh\}$

$\underline{\underline{B}} = \{ad,\ bc,\ ef,\ gh\}$

The first intersection C_{11} is simply _a_; however, _a_ is clustered with _b_ in one case and with _d_ in the second case; hence there is a contradiction which can be removed by deleting both _b_ and _d_. Once these two items are deleted no further contradictions arise and $S_{AB} = 6/8 = .75$.

The maximum value of S_{AB} is 1. The expected value under the hypothesis of no relation between two clusterings depends upon the number of clusters and the size of the clusters in the two analyses. Nothing is known about the sampling distribution of this statistic, but a few Monte Carlo calculationshave provided measures of S_{AB} (not expected values) up to .37 for $S_{\text{Sort 1 - Sort 2}}$.

Table 2 compares the clusterings presented in Table 1 in terms of the statistics developed above. Table 2a shows the proportion of words clustered by each analysis, i.e. $\frac{N - n(X)}{N}$. Except for _Memory 1_ these values are quite high. Note, however, that by the fifth trial organizational effects in the free recall data are just as strong as in the other tasks studied here: 90% of the items are clustered in the _Memory 5_ analysis. Words for which no data were obtained in the _20-question_ game are not included in the figure shown in Table 2a. Table 2b presents the results of the comparisons of the various analyses: the number of words clustered by both, the number of words clustered alike, and the similarity coefficients. The uniformly high values of the latter are apparent.

The high similarity coefficients for the _Memory 1_ analysis should be especially noted: order of output on the first free recall trial was not structured very strongly in that only little more than half of always presented in the same order. Ss were asked to sort the 40 randomly arranged words of each list into from two to eight categories "according to the meaning of the words". A board with eight bins marked on it was in front of S. The task was self-paced. As soon as the deck was sorted S was asked to resort it once, making any changes that seemed necessary.

Results

On the basis of the sorting data proximity matrices were constructed for each list, separately for Sort 1 and Sort 2. The proportion of times that two items were sorted into the same bin was taken as a measure of the proximity between the two items. Thus, a symmetric matrix was obtained with all diagonal elements equal to one.

Proximity matrices for the recall data were constructed similarly. The proximity between two words was estimated by how often the two words were recalled next to each other. More specifically, the entry in the i^{th} row and j^{th} column was one-half times the sum of the proportion of times that item j followed item i in recall, out of the total number of times that item i was recalled, and the proportion of times item i followed item j when j was recalled. This procedure generates symmetric proximity matrices.[4] All diagonal elements were 1, because an item is always recalled with itself.

Both sets of proximity matrices were analyzed as in Experiment I. The proportion of items clustered, as well as the similarity indices for the comparisons between the two sorts and free recall and sorting are shown in Table 3. The similarity indices were about as high as those observed in Experiment I, thus fully confirming our other results. The only difference between the two experiments is the much higher proportions of items clustered in the memory data of Experiment II: while only .55 of the items were clustered in the data from the first recall trial in Experiment I, a corresponding average proportion of .89 was observed in Experiment II. This may perhaps reflect differences in list structure, but it may also be due to the fact that Ss in Experiment I worked at a very slow pace.

Since the clusters obtained from the sorting data and the clusters obtained from the memory data overlapped to a large extent for all six lists, we conclude that there is at least some amount of generality over different lists for the phenomena described in Experiment I.

Discussion

Word lists that are structured in some way are easier to recall than comparable unstructured lists, and related words tend to be recalled in clusters. Somehow words that are organized in a stable structure are more accessible in memory. The present study was designed as an initial investigation of the mechanism that is responsible for this phenomenon. It is not concerned with why organized material is easier to recall, nor how the recall process works; it merely demonstrates that recall involves in some important way the structure of the word lexicon that is part of a person's long-term memory. This much one can conclude from the observation reported here that the output order in free recall is structured in much the same way as performance in other tasks that depend upon how memory is organized. In some as yet unspecified way the manner in which information about words is coded in S's memory determines clustering in free recall. Since the improved recall of organized lists is closely related to the formation of higher-order memory units, and hence to clustering, one may conclude that one important class of recall cues is provided by the relationships among lexical items in S's long-term memory. How lexical structure could have such an effect was discussed by Kintsch (1970b).

Throughout this paper the terms "structure of memory" and, more specifically, "lexical structure" have been used. It has not been our purpose to present a detailed theory of the organization of memory, nor is such a theory a prerequisite for the understanding of the work reported here: whatever the organization of memory is, it places similar constraints upon performance in the tasks investigated here, and the only purpose of the present series of experiments was to determine how great this similarity is. Presumably the organization of memory involves many kinds of semantic relationships as well as imagery.[5]

That input order can be a very powerful determinant of clustering in free recall has been noted before. There exists also a very interesting interaction between input order and semantic structure: semantically related items that are close together in input are much more likely to be clustered together than related items that are not adjacent in input. In order for semantic relationships to become effective they must be noticed by S: what seems to happen is that words that are held together in a short-term memory buffer are scanned for the presence of semantic relationships among them, while the semantic relationships among words that do not occupy short-term memory at the same time may be missed. Glanzer (1969) found that associated words are more likely to be recalled if the number of intervening input words is small. Similarly, Kintsch (1970b) showed that clustering in free recall is a function of the input distance between related items. On the basis of a cueing study Wood (1969) also concluded that more higher-order memory units are formed than can be retrieved when related words are presented consecutively, but not when related items are separated in input. Note, however, that input adjacency is not an absolute requirement for clustering: through implicit recall of items ("rehearsal", "mediating responses") related words that appeared in different parts of the input list may be brought to the short-term working memory, as was emphasized by both Kintsch (1970b) and Wallace (1970). In addition, both of these authors note that free recall learning is not merely a matter of exploiting preexisting relations among items but may consist in establishing new relationships.

A final comment needs to be made about the use of cluster analysis in the present paper, to avoid possible misunderstandings on this score. Normally in a cluster analysis one is interested in what clusters there are and what the properties of the cluster space are, and one wants to make sure that the obtained relationships are both reliable statistically and interpretable. In the present application it was considered irrelevant what the "semantic space" is like, or how reliable the results are statistically. The analysis was used here simply to discover clusters in order to be able to compare these clusters with each other, as long as these clusters were obtained by the same procedure, based upon the same criteria. Of course, if the clusterings were either nonsensical or grossly unreliable, one would hardly obtain as much overlap as has been found here.

References

Battig, W. F., Allen, M., and Jensen, A. R., Priority of free recall of newly learned items. _Journal of Verbal Learning and Verbal Behavior_, 1965, _4_, 175-179.

Cofer, C. N., Segal, E., Stein, J., and Walker, H. Studies on free recall of nouns following presentation under adjectival modification. _Journal of Experimental Psychology_, 1969, _79_, 254-264.

Glanzer, M. Distance between related words in free recall: trace of the STS. _Journal of Verbal Learning and Verbal Behavior_, 1969, _8_, 105-111.

Kintsch, W. _Learning, memory, and conceptual processes_. New York: Wiley, 1970. (a)

Kintsch, W. Models for free recall and recognition. In D. A. Norman, _Models of Human Memory_. New York: Academic Press, 1970. Pp. 331-374. (b)

Mandler, G. and Dean, P. J. Seriation: development of serial order in free recall. _Journal of Experimental Psychology_, 1969, _81_, 207-215.

Mandler, G. and Griffith, N. Acquisition and organization of new items in the free recall of random lists. _Journal of Verbal Learning and Verbal Behavior_, 1969, _8_, 545-551.

Miller, G. A. A psychological method to investigate verbal concepts. _Journal of Mathematical Psychology_, 1969, _6_, 169-191.

Riegel, K. F. The Michigan restricted association norms. Technical Report No. 3, 1965, University of Michigan.

Roberts, W. A. The priority of recall of new items in transfer from part-list learning to whole-list learning. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 645-652.

Stanners, R. F. Grammatical organization in free recall. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 95-100.

Tryon, R. C. and Bailey, D. E., Cluster Analysis. New York: McGraw-Hill, 1970.

Wallace, W. P. Consistency of emission order in free recall. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 58-68.

Wood, G. Retrieval cues and the accessibility of higher-order memory units in multi-trial free recall. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 782-789.

Footnotes

1. This research was supported by grant MH-15872 from the National Institute of Mental Health. We thank Daniel E. Bailey for making available his computer program for the cluster analyses, and Klaus F. Riegel for the use of the Michigan Restricted Association Norms.

2. A second calculation of overlap scores was performed in which the actual frequencies with which responses were given were used. For instance, 10 Ss gave SKY as a response to BUTTERFLY and 30 Ss gave SKY as a response to BIRD in the example cited above; hence an overlap score of 10 was recorded. However, this way of calculating overlap scores was not worth the effort: the data clustered in almost the same way as in the simpler analysis.

3. Whenever the BC TRY analysis produced a cluster in which some items had positive and some items had negative factor loadings, it was divided into a negative and into a positive cluster, because negative loadings imply that the items are opposite geometrically to the items with positive loadings.

4. Symmetric proximity matrices were used here, unlike in Experiment I, because several different methods of analysis were tried with the data, before we determined that the BC TRY analysis was the most satisfactory one for present purposes. (Experiment II was conducted before Experiment I.)

5. Wallace (1970) proposes an associative theory of clustering in free recall. However, calling all relationships that are known to be important determinants of clustering "associative" broadens the use of the term much beyond its operational definition, and a more neutral designation would seem preferable.

Table 1

Results of the cluster analyses for Sort 1, Sort 2, Association, Twenty-Questions, and Trials 1 and 5 of Free Recall. For explanation see text.

		BIRD	BUTTERFLY	INSECT	SHEEP	SPARROW	BABY	DOCTOR	GIRL	NURSE	PROFESSIONAL	ARM	BRAIN	HAND	MOUTH	STOMACH	BUTTER	CHEESE	FOOD	MILK	BED	DESK	CHAIR	FURNITURE	TABLE	BARN	BUILDING	HOUSE	ROOF	WINDOW	ANGER	COMFORT	HATE	JOY	TROUBLE	COMMAND	JUSTICE	LEADERSHIP	SICKNESS	SEX
SORT 1	a																				X	X	X	X	X	X	X	X	X	X										
	b																X	X	X	X																				
	c											X	X	X	X	X																								
	d	X	X	X	X	X																																		
	e							X		X	X																													
	f						X		X																															
	g																														X	X	X	X	X	X	X	X	X	
SORT 2	a																				X	X	X	X	X	X	X	X	X	X										
	b																X	X	X	X																				
	c											X	X	X	X	X																								
	d	X	X	X	X	X																																		
	e						X	X	X	X	X																													X
	f																														X	X	X	X	X	X	X	X	X	
ASSOC.	a																							X			X	X	X	X										
	b																X	X	X	X																				
	c											X	X	X																										
	d	X	X	X		X																																		
	e																														X	X	X	X						
	f																																		X		X			
	g																																			X		X		
	h							X		X	X																												X	
	i						X		X														X																	
QUEST.	a																				X	X	X	X	X	X	X	X	X	X										
	b	X	X	X	X													X		X																				
	c						X	X	X	X	X	X	X	X	X	X																								
	d																														X		X	X					X	
	f																															o			o	o	o	o		o
MEM 1	a																								X			X												
	b																							X			X		X											
	c																X	X		X																				
	d											X		X	X	X																								
	e	X				X																																		
	f		X	X																																				
	g							X		X																														
	h						X		X																															X
MEM 5	a																									X	X	X	X	X										
	b																					X	X		X															
	c								X															X																
	d																X		X	X																				
	e											X		X		X																								
	f	X			X	X																																		
	g		X	X																																				
	h																														X	X	X	X						
	i																																			X	X	X		
	j							X		X	X																												X	
	k						X														X																			X

Table 2

(a) Proposition of words clustered for the different tasks in Experiment 1:

Sort 1	Sort 2	Associations	Questions	Memory 1	Memory 5
.97	1.00	.80	.91	.54	.90

(b) Comparison between analyses:

	Number of words clustered by both	Number of words clustered alike	S_{AB}
Sort 1 - Sort 2	38	38	1.00
Sort 1 - Associations	31	29	.94
Sort 1 - Questions	30	30	1.00
Sort 1 - Memory 1	20	20	1.00
Sort 1 - Memory 5	34	31	.91
Sort 2 - Associations	31	29	.94
Sort 2 - Questions	30	30	1.00
Sort 2 - Memory 1	21	21	1.00
Sort 2 - Memory 5	35	32	.91
Associations - Questions	23	21	.91
Associations - Memory 1	17	17	1.00
Associations - Memory 5	28	27	.96
Questions - Memory 1	18	18	1.00
Questions - Memory 5	27	24	.88
Memory 1 - Memory 5	19	17	.89

Table 3

(a) Proportion of items clustered for the six different word lists in Experiment II:

List:	1	2	3	4	5	6	Average
Memory:	.83	.90	.70	.72	.93	.90	.89
Sort 1:	.83	.98	.92	.78	.98	.92	.92
Sort 2:	.85	.98	.88	.93	.93	.90	.91

(b) Similarity indices S_{AB} for the same six word lists:

Sort 1 - Sort 2	.96	1.00	.8[illegible]	1.00	1.00	1.00	.98
Sort 1 - Memory	.85	.91	.88	1.00	.94	.79	.90
Sort 2 - Memory	.83	.91	.96	.93	1.00	.94	.93

Kim Romney

Toward a theory of Analogical Reasoning

by

David E. Rumelhart

and

Adele A. Abrahamson

University of California, San Diego

This paper was prepared for a Research Workshop on Cognitive Organizations and Psychological Processes sponsored by the Committee on Basic Research in Education National Research Council. This draft is as yet incomplete. We invite your critical comments.

Toward a Theory of Analogical Reasoning[1]

David E. Rumelhart

and

Adele A. Abrahamson

University of California, San Diego

Despite psychologists considerable confidence that analogical reasoning plays an important role in intelligent behavior (c.f. Bartlett, 193 ; Miller, 194 ; Minsky, 1966; Oppenheimer, 195 ; and Polya, 1945; Reitman, 1964) analogy formation has received little systematic attention. Nearly all of the work which has been done has been in the form of complex computer programs (c.f. Becker, 1969; Evans, 1968; Reitman, 1964). No systematic comparisons of the programs with subject's behavior have been carried out (see Hunt, 1968). The present paper outlines a simple theoretical model for the understanding of analogical reasoning and evaluates the validity of the model by means of empirical test.

To introduce our notion of analogical reasoning it is useful to outline a definition of the word reasoning from which we can work. The term is used here to denote those processes in information retrieval which depend on the structure, as opposed to the content of organized memory. Thus, one might answer the question "Who is the father of your country?" in at least two different ways. In one case, the specific information that George Washington was the "father of our country" might be stored and used to answer the question. On the other hand, when specific information is not available one can consult the stored meanings of the words in question and one's knowledge of history to derive a plausable answer. The first of these methods might be called remembering since retrieval depends primarily on the specific information stored. The second method may be identified with reasoning, since in this case retrieval depends to a much greater extent on the form of the relationship among the words. The same act of reasoning (i.e., the same processes) could have been applied to the question "Who was the father of your state?" or "Who was the mother of your country?". It is not the specific content of the question but the form of the relationships among the words which determines the response.

If one accepts this working definition of reasoning, then the theoretical problem in understanding any particular reasoning process becomes clear. We must (1) specify the form of the memory structure and then (2) determine the algorithm which is applied in the case of the reasoning process in question.

Perhaps the simplest reasoning task by our definition involves the judgment of similarity or dissimilarity of concepts. It is normally assumed that similarity is not directly stored and thus "remembered" as such. Instead, it is assumed ~~rather~~ that it is derived from the memory structure. The simplest view holds that similarity between concepts is a simple function of the "psychological distance" between these concepts in the memory structure. The "closer" two concepts are to one another in memory, the more similar they are. Thus in the case of judging similarities the two questions which we need to answer are: (1) what is the nature of the memory structure which underlies similarity judgments; and (2) what is the measure of "distance" on this psychological space?

A recent paper by Nancy Henley (1969) illustrates one set of answers which have been given to these questions. Henley assumed (1) that the memory

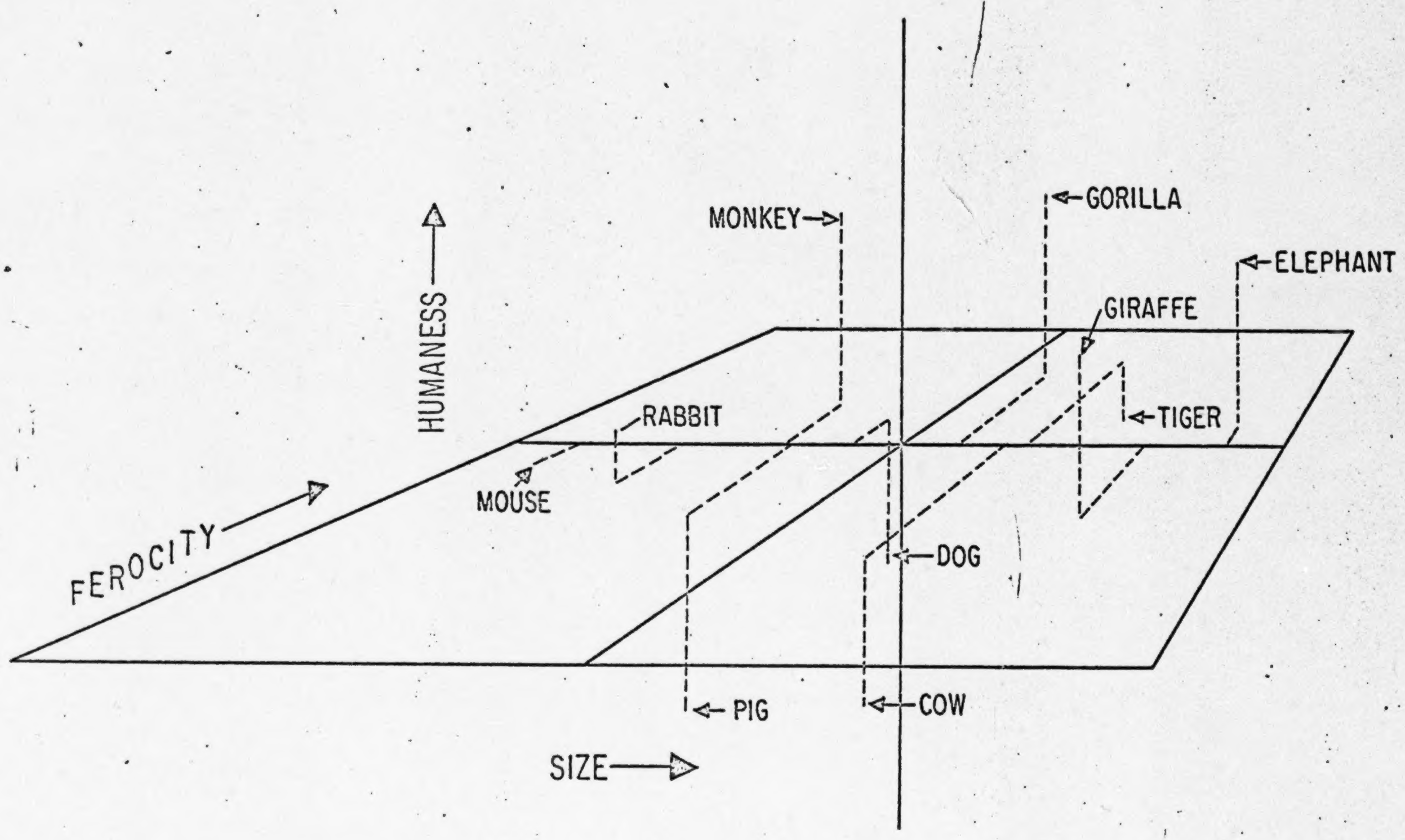


Figure 1. The placements of a selected set of animals from Henley (1969).

structure may be represented as a multidimensional Euclidean space and (2) that similarity is inversely related to distance in this multidimensional space. To test these assumptions in the semantic domain Henley used the techniques developed by Shepard, (1962a,b) Kruskal, (1964a,b) and others to deduce the form of the space from subjects judgments of similarity and dissimilarity among the concepts.

In one of her experiments Henley used ratings of dissimilarity to deduce the underlying "psychological space" relating common animal terms to one another. She chose thirty of the most common mammals and asked subjects to rate all possible pairs of the thirty mammals as to their dissimilarity on a scale from 0 to 10. A value of 0 indicated that the animals were identical, a value of 10 indicated that they were maximally different. These data were used as input to TORSCA, (a multidimensional scaling program developed by Young and Torgerson; 1967) to find the multidimensional solution with minimum dimensionality consistent with the observed dissimilarity data.

Henley's Her results showed that these thirty mammals fitted reasonably well into a three dimensional space (Kruskal's stress index at 9.4%). Figure 1 shows

Insert Figure 1 about here

the three dimensions along with placements of some on the animals. Henley employed several other methods including the method of triads to derive the semantic space and obtained remarkably similar results in each case.

This example illustrates one line of work on the problem of specifying the form of the memory structure and the method of deriving similarity judgments from the structure. Most of the work to be discussed in this paper depends on a similar set of assumptions. If we accept the view that at least portions of structure of semantic memory can be represented as a multidimensional space then in order to specify any particular reasoning process operating on this structure we need only specify the algorithm associated with this reasoning process. The case of similarity judgments is a particularly simple one because distance is a particularly simple computation within the frame work of a multidimensional Euclidean space. A little thought will show that analogical reasoning can itself be considered a kind of similarity judgment in which not only the magnitude of the distance but also the direction must be considered. Consider for example the assertion A is to B as C is to D, the classic analogic paradigm. When we make this assertion we actually mean that the concept A is similar to B in exactly the same way and to exactly the same degree as C is similar to D. That is, in the multidimensional representation, we are asserting the directed or vector distance between A and B is exactly the same as the directed or vector distance between C and D. Thus, in an analogy problem of the Form A is to B as C is to what, the proper answer must be that concept which is most nearly the same vector distance from C as B is from A. These ideas are stated more formally and specifically in the set of assumptions listed below.

Consider an analogy problem of the form $A:B::C:(X_1,X_2,\ldots,X_n)$. (To be read A is to B as C is to which of the following: $X_1,X_2,\ldots,$ or X_n). It is assumed that:

A1. Corresponding to each element of the analogy problem there is a point in an m-dimensional space. (We denote, for example, the point corresponding to element A of the problem as $\underline{A}$ and say that the coordinates of $\underline{A}$ are the ordered sequence $\{a_j\}_{j=1,m}$, where a_j is the coordinate value of $\underline{A}$ on dimension j.)

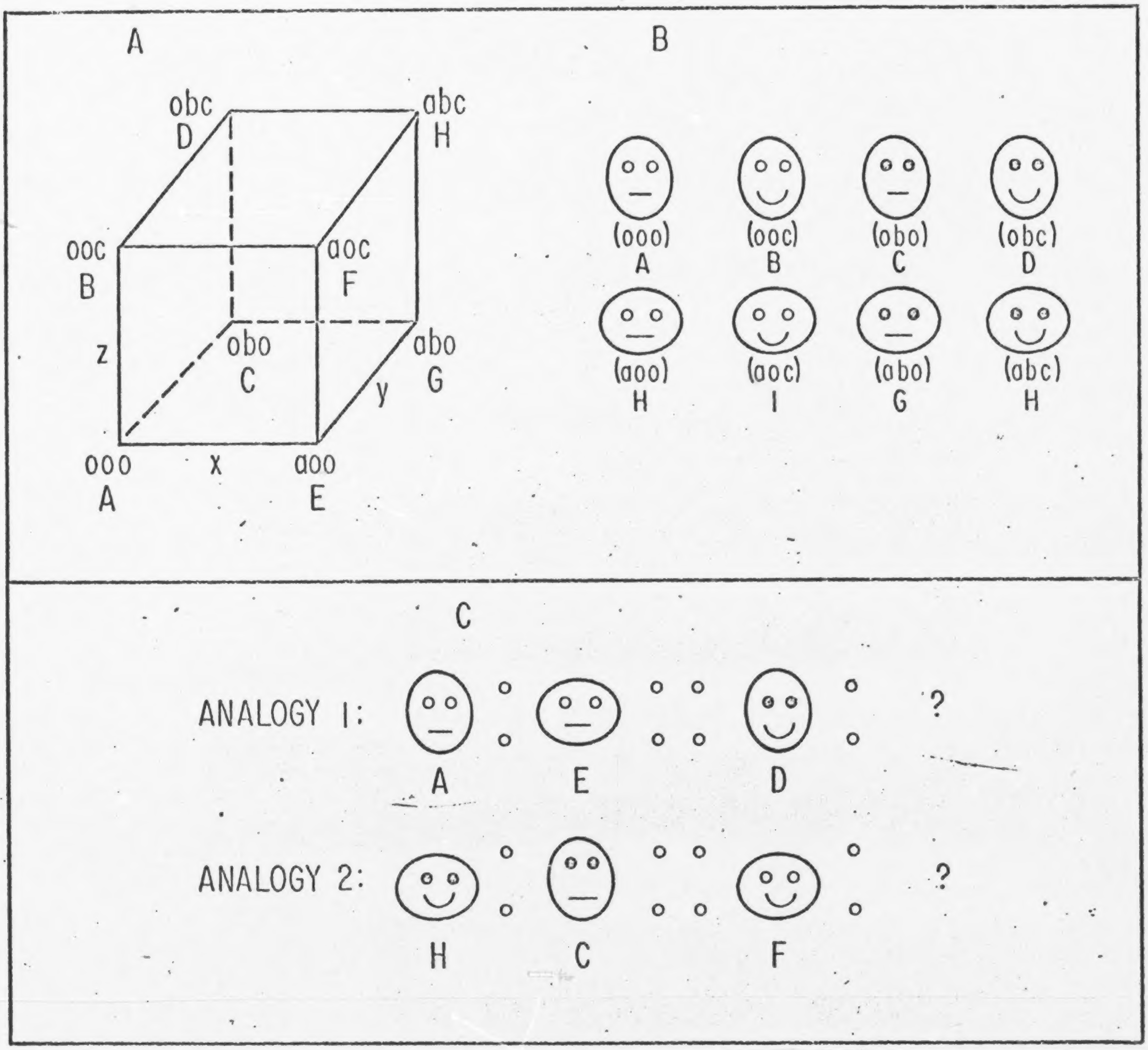


Figure 2. Panels (A) and (B): Three dimensional representation of similarities among eight faces. (Example taken fro Tversky and Krantz; 1969). Panel (C): Two sample analogies generated from the faces.

A2. For any analogy problem of the form A:B::C:? there exists a concept I (probably without a name) such that A:B::C:I and an ideal analogy point, denoted $\underline{I}$ such that $\underline{I}$ is located the same vector distance from $\underline{C}$ as $\underline{B}$ is from $\underline{A}$. The coordinates of $\underline{I}$ are given by the ordered sequence $\{c_j+b_j-a_j\}_{j=1,m}$.

A3. The probability that any given alternative X_i is chosen as the best analogy solution from the set of alternatives $X_1,\ldots,X_n$ is a monotonic decreasing function of the absolute value of the distance between the point $\underline{X}_i$ and the point $\underline{I}$ denoted $|\underline{X}_i-\underline{I}|$.

To summarize, we assume that each element in an analogy can be represented as a point in an m-dimensional Eucledian space, that the ideal analogy solution is given by that point in the space which lies the same vector distance from C as B lies from A, and that, the closer a given alternative is to the ideal analogy solution, the higher the probability it will be chosen as the best analogy.

The intuitions behind these assumptions can be further illustrated with reference to the eight faces in Figure 2. Each face corresponds to a corner of the three-dimensional similarity space shown in the figure. The coordinates of each face are shown beside the face. Consider, as an example analogy of the form: A:E::D: (B,C,F,G,H). For simplicity assume that the space shown in Figure 2 satisfies A1. Then from A2 we get the coordinates of $\underline{I}$ to be

----

Insert Figure 2 about here

$$\{d_j+e_j-a_j\}_{j=1,3} = (o+a-o,\ o+b-o,\ c+o-o) = (a,b,c)$$

Hence, we conclude that the face corresponding to the ideal solution for this analogy is face H, with coordinates (a,b,c). Thus from A3 we conclude that face H should be the most probably chosen face from the set.

The second analogy illustrated in the figure is similar $H:\overset{C}{\cancel{B}}::\overset{F}{\cancel{G}}:(A,\overset{B}{\cancel{C}},D,E,\overset{G}{\cancel{F}})$.

Here face A with coordinate (o,o,o) corresponds to the ideal analogy solution. Obviously numerous other similar examples could be generated from this set of stimuli, but we turn now to a more serious test of the assumptions.

Experiment I[2]

Assumption A1 of the theory requires a set of concepts located in a multidimensional space. The set chosen, and used throughout Experiments I, II and III is the animal set scaled by Henley (1969) and described above. This set was chosen because the space seemed robust ~~(Henley, 1968)~~ and because college student subjects are all generally familiar with the set of words. All analogy problems in Experiment I were of the familiar form A:B::C: (D_1,D_2,D_3,D_4).

Procedure

Analogy problems were generated in the following way. Animals from the set of thirty were chosen at random without replacement. The first animal chosen was the first element of the first analogy problem. The second chosen was the second element of the first analogy problem; the third chosen was the third element of the first problem. The fourth chosen was the first element of the second analogy problem etc. This procedure was continued until the set of animals was exhausted (i.e., ten analogies were formed). The animals were then reshuffled and the procedure was repeated until thirty unique analogy problems were formed. For each of the thirty analogy problems four alternatives were chosen. One alternative was chosen, at random, from among those animals within .5 units of the ideal analogy solution (.5 units corresponds roughly to the distance between camel and antelope. If no animals fell within .5 units the analogy was discarded and a new analogy was formed. After the first alternative was chosen a second was chosen from among those animals between .5 and 1.0 units from the ideal solution. (One unit corresponds to the distance between deer and elephant). The third alternative was chosen among those animals from between 1.0 and 1.5 units from the ideal solution. (A distance of 1.5 units corresponds to the distance between elephant and pig.) The fourth and final alternative was chosen from among those animals more than 1.5 units from the ideal solution.

Thirty-five subjects were recruited from a lower division psychology class at the University of California, San Diego. Subjects were run in two groups, one of 15 and one of 20 students. Each subject was given a five page mimeographed booklet. The first page contained the instructions and an example. Subsequent pages contained the thirty analogy problems. The four response alternatives were listed in ~~random~~ alphabetic order. Examples of the analogy problems are given in Table 1. The directions were as follows:

Insert Table 1 about here

DIRECTIONS:

An analogy task is one in which you recognize relationships of things or ideas to other things or ideas. In this task you will be given, the last term in the analogy will be missing. You will have four choices from which you will supply the term which seems to you to be the most analogous. Then you should indicate your second, third, and fourth choices, also. After that, you should indicate which of the terms you could consider to be analogous and which ones you think are not analogous by drawing a cut-off line between the two "groups."

Example: apple:tree::grape: __________

A. bush
B. vine
C. barrel
D. ground

1. B
2. D
3. A
4. C

Table 2

Subjects' rankings as a function of alternative distance.

		Observed Rank			
		1	2	3	4
Rank Distance of the Alternative	1	.709	.180	.069	.046
	2	.177	.546	.137	.129
	3	.086	.160	.526	.226
	4	.043	.111	.243	.600

> The ordering of the relationship in an analogy is important. For instance, in the above example apple:tree:vine: grape would be incorrect because the relationship of apple to tree is not the same as that of vine to grape.
>
> All of the terms in the following analogies are animal names. Feel free to go back and change any answer.

Results and Discussion

The basic results of Experiment I are shown in Table 2. Table 2 gives

---

Insert Table 2 about here

---

the proportion of responses, averaged over subjects and analogy problems, for which Rank 1 was given to the jth closest response alternative. Thus the upper left hand entry of the table implies that 70.9 percent of the Rank 1 responses were given to the response alternative closest to the ideal point. Only 4.6 percent of the Rank 1 responses were given to the most distant response alternative.

Although the theory outlined in assumptions A1-A3 made no specific predictions beyond the Rank 1 data it is clear that the entire table is consistent with the ideas behind the theory. As it stands assumption A3 asserts only that a monotonic decrease in probability should be observed in the first column of Table 2. Before more specific predictions can be made we require a more specific decision rule. The rule we propose is that developed by Luce (1959). We thus substitute for assumption A3 assumptions A3' and A4 outlined below.

A3'. The probability that any given alternative X_i is chosen from the set of alternatives $X_1,\ldots,X_n$ is given by

$$\Pr(X_i|X_1,\ldots,X_n) = p_i = \frac{V(d_i)}{\sum_{j=1}^{n} V(d_j)} ,$$

where $d_i = |\underline{X}_i - \underline{I}|$ denotes the absolute value of the distance between $\underline{X}_i$ and $\underline{I}$ and $v(i)$ is a monotonically decreasing function of its argument.

A4. $v(x) = \exp(-\alpha x)$ where x and α are positive numbers.

Thus, A3' is simply a restatement of Luce's choice rule and A4 is an assertion that the monotonically decreasing function of A3' is an exponential. The exponential was chosen for a number of reasons. It was chosen in part because Shepard (1957) found a good fit to an exponential generalization function over a "psychological space" derived in a similar fashion, and in part because it is a simple one parameter decay function. Assumptions A3' and A4 were stated separately because of the obvious possibility that A3' be correct and A4 false. Stated in this way it is easy to separate the Luce choice rule and the exponential decay assumption.

Figure 3 is a plot of the predicted versus the observed number of subjects ranking each of the 4 alternatives of each of the 30 analogy problems as the best solution to the analogy. The single parameter of the fit, α, was estimated by a least squares procedure to be 2.9. A product moment correlation computed from these data show a correlation of .933 between the predicted and observed values.

Insert Figure 3 about here

We can thus conclude, that the set of assumptions A1,A2,A3' and A4 yield an adequate account of the first choice data. It is furthermore clear from the orderly nature of the data in Table 2 that a slight extension of the theory to include a theory of rankings may well account quantitatively for the entire matrix of data displayed in the table. Given our use of Luce's choice axiom in A3' it is natural to accept Luce's extension of his choice

Figure 3. Predicted versus observed number of subjects ranking each alternative as the best analogy solution.

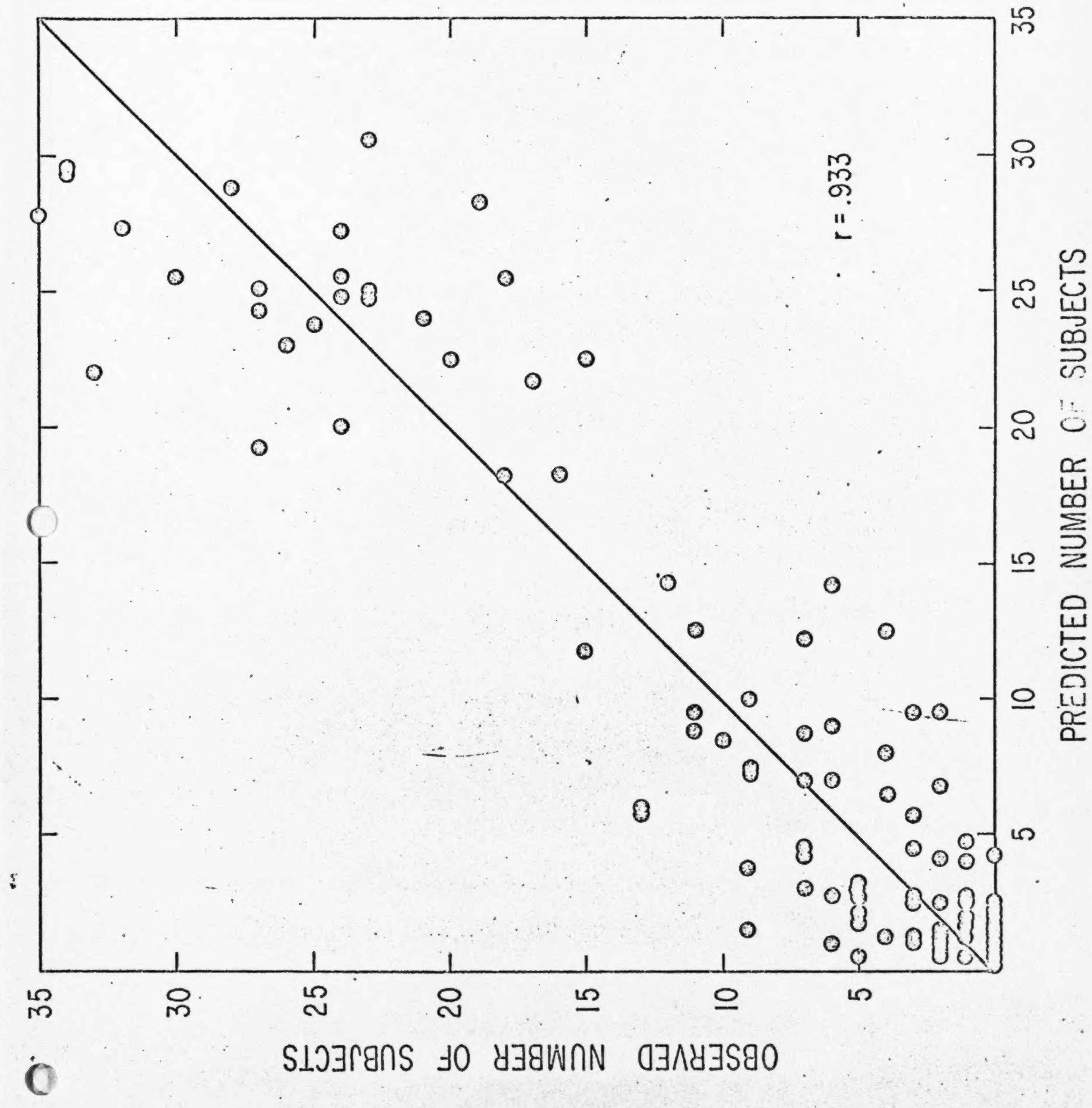

r = .933
PREDICTED NUMBER OF SUBJECTS
OBSERVED NUMBER OF SUBJECTS
5
10
15
20
25
30
35
5
10
15
20
25
30
35

Table 1

Example analogies and solutions from Experiment I.

RAT:PIG::GOAT:______

A. CHIMPANZEE
BB. COW
C. RABBIT
D. SHEEP

RANKS
1. B
2. D
3. C
4. A

CAMEL:DONKEY::RABBIT:______

A. ANTELOPE
B. BEAVER
CC. CAT
D. TIGER

RANKS
1. B
2. C
3. A
4. D

FOX:HORSE::CHIPMUNK:______

A. ANTELOPE
B. DONKEY
C. ELEPHANT
D. WOLF

RANKS
1. C
2. A
3. B
4. D

LION:WOLF::GOAT:______

A. CAT
B. CHIMPANZEE
C. GORILLA
D. PIG

RANKS
1. D
2. A
3. B
4. C

axiom to cover rankings. We thus append an additional assumption, A5, to our list.

A5. We assume that the subjects rank a set of alternatives by first choosing the Rank 1 element according to A3' and, then, of the remaining alternatives deciding which is superior by application of A3' to the remaining set and assigning that Rank 2. This procedure is assumed to continue until all alternatives are ranked.

Figure 4 is plot of the predicted and observed mean number of subjects,

---

Insert Figure 4 about here

---

averaged over analogy problems, to assign rank j to the ith closest response alternative, for Ranks 1 through 4. The value of $\alpha = 2.9$ used to generate the theoretical curves was taken from fit of the first choice data discussed above. No new estimate was made to fit the 2nd, 3rd and 4th ranked alternatives.

Although the results of Experiment I are certainly encouraging, it is somewhat unfortunate that the design of the experiment required an extension of the original three assumptions into five before quantitative tests of the theory could be made. The second experiment was designed to get a test of the basic theory unencumbered by the added response assumptions.

Experiment II

The goal of this second experiment was to extend the results of Experiment I by finding evidence bearing more directly on the basic assumptions of the model. The strong implication of assumptions A1-A3 is that the probability of choosing any particular alternative X_i as the best alternative depends only on the ideal solution point I and on the alternative set $(X_1,\ldots,X_n)$ but not at all on the analogy itself. Thus, all possible analogies with a given

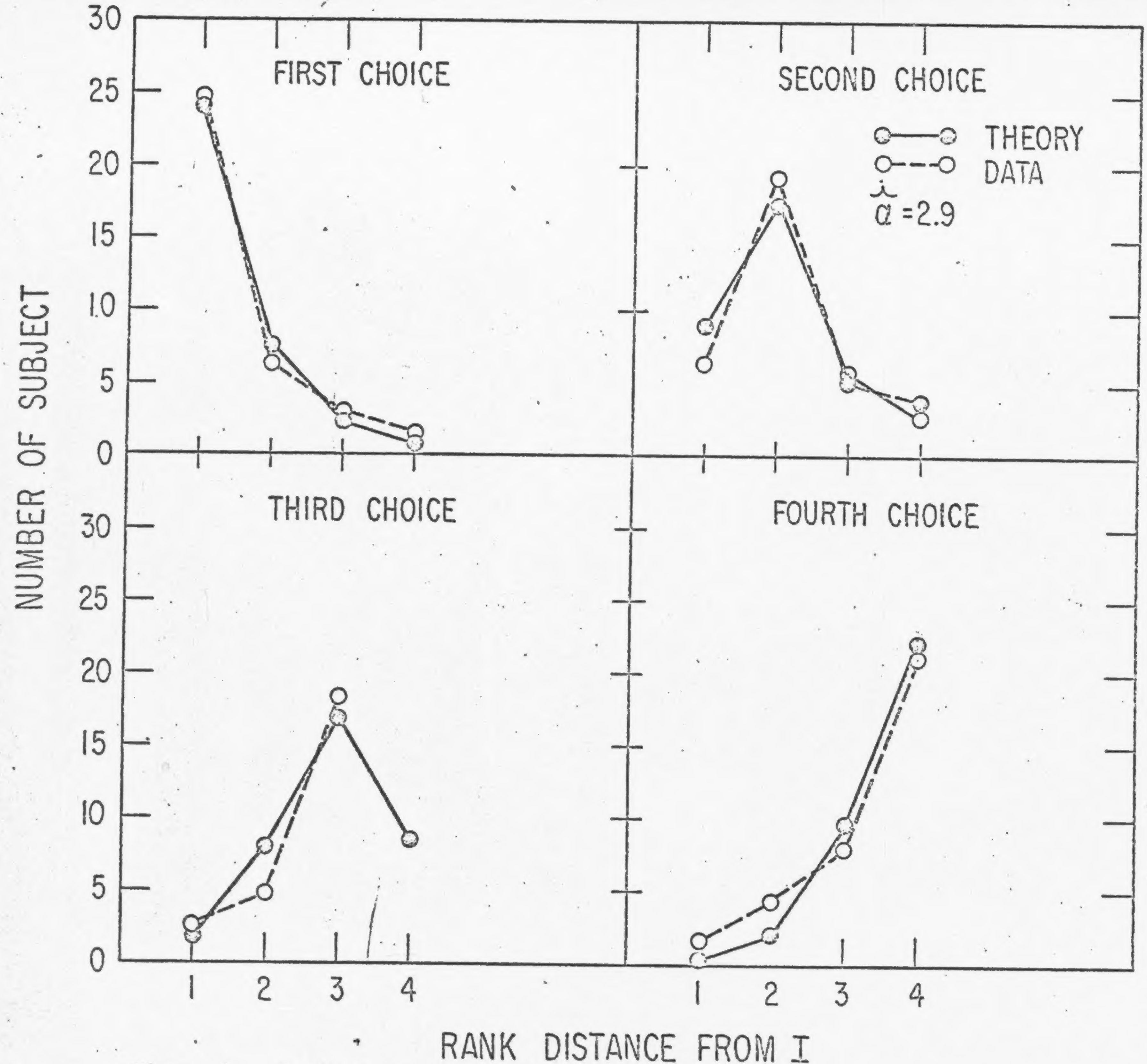


Figure 4. Predicted versus observed mean number of subjects ranking each of

ideal solution point I and a given alternative set ($X_1, \ldots, X_n$) should yield the same distribution of responses over the X_i. Experiment II was designed primarily to test this implication.

Procedure

Twelve pairs of analogy problems were constructed such that each pair had the same ideal analogy point (within a tolerance of ~~.03~~ .12 units; roughly the distance between lion and tiger. For each analogy pair two sets of alternatives (denoted set A and set B) were constructed. The alternative sets were constructed with the constraints that (1) there be no overlap between them; (2) the ith closest alternative for one set is roughly the same distant as the ith closest alternative for the other set, ~~i = 1 to 4~~ for all i; and (3) alternatives of a given set must be roughly equally separated in distance from one another by either .2,.25,.35,.4 or .5 units.

The forty-four subjects were divided into two groups. Each group was given 24 analogy problems to solve. Each member of each of the twelve analogy pairs were given to each subject. However, no subject received the same alternative set more than once. Half of the subjects received alternative set A for the first analogy of each pair and set B for the second. The other half received alternative set B for the first analogy of the pair and set A for the second.

The twenty-four analogy problems were presented in booklet form. The instructions followed closely those given in Experiment I. The analogy problems appeared in random order and the alternatives for each analogy were also scrambled. Examples of two of the analogy pairs along with their alternative sets are given in Table 3.

Table 3

Two analogy pairs with alternative sets from Experiment II.

ANALOGY PAIR 1	GORILLA:DEER::BEAR:______	BEAVER:SHEEP::DOG:______
ALTERNATIVE SET	A	B
	1. COW	1. DONKEY
	2. PIG	2. CAMEL
	3. TIGER	3. ELEPHANT
	4. MONKEY	4. CHIMPANZEE
ANALOGY PAIR 4	CAT:SHEEP::LEOPARD:______	MOUSE:RACCOON::COW:______
ALTERNATIVE SET	A	B
	1. HORSE	1. ZEBRA
	2. DEER	2. GIRAFFE
	3. FOX	3. LION
	4. RAT	4. CHIPMUNK

Insert Table 3 about here

Results and Discussion

The theory holds that the response distribution should depend only on the ideal solution point and the alternative set, and not on the particular analogy problem. This prediction was tested by computing a X^2 between the response distributions given to the two halves of each analogy pair for each response set. These X^2 values are shown in Table 4. The summed X^2 reaches a value of 109.7, with 60 degrees of freedom. This value is somewhat larger than would be expected by chance if the theory were true. On the other hand a close look at the table will show that most of the deviation is contributed by a very few comparisons. If, for example, the two largest values are ignored, the summed X^2 reduces to 84.0 with 54 degrees of freedom, a relatively probable value. Considering all of the possible sources of error it seems safe to conclude, at least to a first order of approximation, that subjects responses depend on the ideal analogy point and the alternative set, but not on the particular analogy problem. It should be pointed out, however, that in certain isolated examples, such as analogy pair 6, alternative set A, there is quite clearly an interaction between the particular alternatives and the analogy problem itself.

Insert Table 4 about here

Since on the whole assumptions A1 through A3 are confirmed by our analysis it makes sense to analyze these data with respect to the remaining assumptions of the theory. It is possible in the context of Experiment II to

Table 4

Analogy[3] Pair	Alternative Set	X^2	d.f.	Analogy Pair	Alternative Set	X^2	d.f.
1	A	1.83	2	7	A	7.33*	2
	B	.11	4		B	7.42*	2
2	A	.39	2	8	A	1.68	2
	B	.93	2		B	9.47	4
3	A	3.03	4	9	A	6.70*	2
	B	7.66	4		B	10.47**	2
4	A	2.47	2	10	A	.86	2
	B	3.39	4		B	.39	2
5	A	13.50**	2	11	A	1.47	2
	B	---	-		B	.12	2
6	A	25.71**	4	12	A	1.65	2
	B	2.12	4		B	1.00	2

Summed $X^2 = 109.7$ d.f. = 60

* Value exceeds .05 significance level.
** Value exceeds .01 significance level.

get a direct look at the form of $v(x)$ and thus a direct evaluation of assumption A4. To see how this is possible consider assumption A3'. From A3' we have

$$p_i = \frac{V(d_i)}{\sum_{j=1}^{4} V(d_j)}$$

Taking the natural logarithm of each side yields

$$\ln p_i = \ln V(d_j) - \ln (\sum_{j=1}^{4} V(d_j)).$$

Now if $v(x) = \exp(-\alpha x)$ as asserted we get

$$\ln p_i = -\alpha d_i - \ln (\sum_{j=1}^{4} v(d_j)) .$$

In words, we get that the natural logarithm of the probability that alternative i is chosen is a straight line in the distance of $\underline{X}_i$ from $\underline{I}$ with slope $-\alpha$ and intercept $-\ln(\Sigma V(d_j))$. In Experiment II alternatives sets were designed in such a way that several different analogy problems had the same set of interalternative distances thus allowing averaging over analogy problems. There are only five different configurations of interalternative distances. Figure 5 shows the straight line fit to each of the five groups of analogies. The value of α was estimated by a least squares procedure to be 1.68. The fit of the straight line is quite good. The only real deviation is depicted in panel E. This is the condition in which the alternatives are closest together and a condition containing the lowest number of observations.

Insert Figure 5 about here

Although in our analysis to this point we have looked only at the first

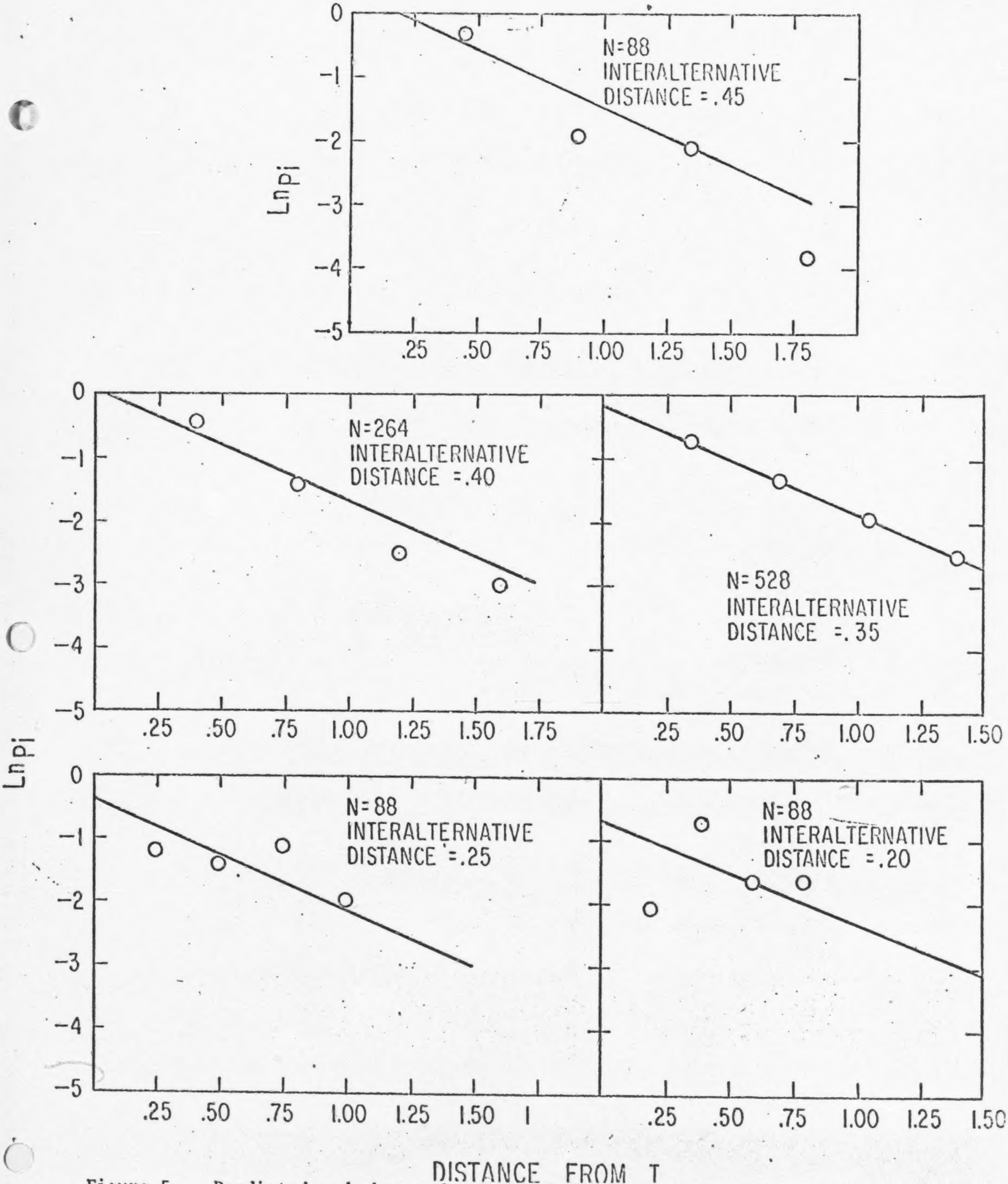


Figure 5. Predicted and observed values of the natural logarithm of the probability that each of the four response alternatives will be chosen as the best analogy solution for each of five sets of

choice data, subjects in Experiment II, as those in Experiment I, were asked to rank the alternative sets. Table 5 shows the predicted and observed mean number of subjects to rank alternative i in position j for each of the five groups of analogy problems. The value $\alpha = 1.68$ was taken from the fit of the first choice data discussed above.

Insert Table 5 about here

In summary then, Experiment II allowed a non-parametric test of assumptions A1 through A3. Although there were some exceptions, on the whole the results supported the model. In addition Experiment II allowed direct investigation of the exponential function in A4. Against the results supported the form of the model. Finally a comparison of the entire rank by alternative distance confusion matrix with the predicted values lent more support to the entire set of assumptions A1 through A5.

Experiment III

This research was originally started as an alternative approach to the problem of concept formation. It was felt that in natural settings much of concept formation, as opposed to instance generalization, the only form studied in the laboratory. Thus, Experiment III was designed to make use of the theory of analogical reasoning outlined above to facilitate the teaching of concepts.

Assumption A2 implies that each statement of the form A:B:C:? implies the existence of some concept I (probably unnamed) against which alternative sets are compared to find the best alternative. The logic of Experiment III was to give a name to this implied concept I and then to see if subjects are

Table 5

Predicted and observed mean number of subjects to rank the ith closest alternative as Rank j as a function of five interalternative distances. Predicted values generated under the assumption $\alpha = 1.68$.

N	Inter-alternative Distance	Rank of alternative distance from I	Rank 1		Rank 2		Rank 3		Rank 4	
			Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.
88	.45	1	13.4	15.8	6.3	3.5	2.0	1.5	.3	1.3
		2	4.8	3.3	8.4	10.8	6.4	7.3	2.4	.8
		3	2.5	2.5	5.8	6.8	8.5	10.5	6.1	2.3
		4	1.3	.5	2.5	1.0	5.1	2.8	13.2	17.8
264	.40	1	11.8	13.8	6.8	4.6	2.8	1.4	.6	2.3
		2	5.7	5.3	7.9	9.5	6.0	4.1	2.6	3.1
		3	2.9	1.8	4.6	4.9	8.0	9.3	6.4	6.0
		4	1.6	1.1	3.7	3.0	5.1	7.2	12.6	10.6
528	.35	1	10.7	10.9	7.0	6.9	3.4	3.0	.9	1.2
		2	6.1	6.0	7.5	8.9	5.9	4.3	2.5	2.8
		3	3.3	3.2	4.7	3.7	7.7	9.3	6.4	5.8
		4	1.9	1.9	2.8	2.5	5.0	5.4	12.2	12.1
88	.25	1	9.8	6.5	6.9	7.5	3.9	6.3	1.4	1.8
		2	6.0	5.5	6.8	4.5	5.8	4.8	3.3	7.3
		3	3.5	7.0	4.7	5.0	6.7	5.0	7.1	5.0
		4	2.6	3.0	3.6	5.0	5.6	6.0	10.2	8.0
88	.20	1	8.5	2.8	6.7	5.8	4.6	11.3	2.2	2.3
		2	6.0	10.5	6.2	4.8	5.8	4.5	4.0	2.3
		3	4.3	4.3	5.1	7.3	6.2	2.8	6.3	8.5
		4	3.2	4.5	4.0	4.5	5.4	3.3	9.4	9.8

able to use this new concept with a new name in the same way that they are able to use concepts that they already know. For example, the experimenter might assert that A:B:C:GOX where GOX is the concept to be learned. He would then ask the subject to use GOX in other cognitive tasks to see if the subject is able to manipulate it in the way he manipulates other concepts. For example, the subject might be asked to solve analogies of the form A:B::GOX: $(X_1,\ldots,X_n)$, or to judge the similarities between GOXes and other animals, or to give a verbal description of a GOX. In short, we would expect that if the subject had actually "understood" the analogy A:B:C:GOX he would be able to do anything with the concept GOX that he could with any other concept -- he would have formed a new concept.

Procedure

Three points were chosen in Henley's animal space. These points were labeled BOF, DAX and ZUK and represented the three concepts that the subjects were to learn. The points were chosen such that they occupied fairly remote parts of the space. Figure 6 is a three dimensional representation depicting the location of the three artificial ~~pseudo~~ animals, BOF, DAX and ZUK relative to the thirty mammals scaled by Henley. Thus, a BOF lies between elephant and camel, a DAX near chimpanzee and a ZUK between fox and wolf.

Insert Figure 6 about here

An anticipation method of teaching was employed. This is, subjects were first given an analogy problem of the form A:B::GOX: $(X_1,\ldots,X_4)$ and then were asked to make ~~this best~~ a guess as to the best alternative. Following their guess subjects were informed as to the correct alternative along with a ranking of the remaining alternatives from best to worst. After a period of

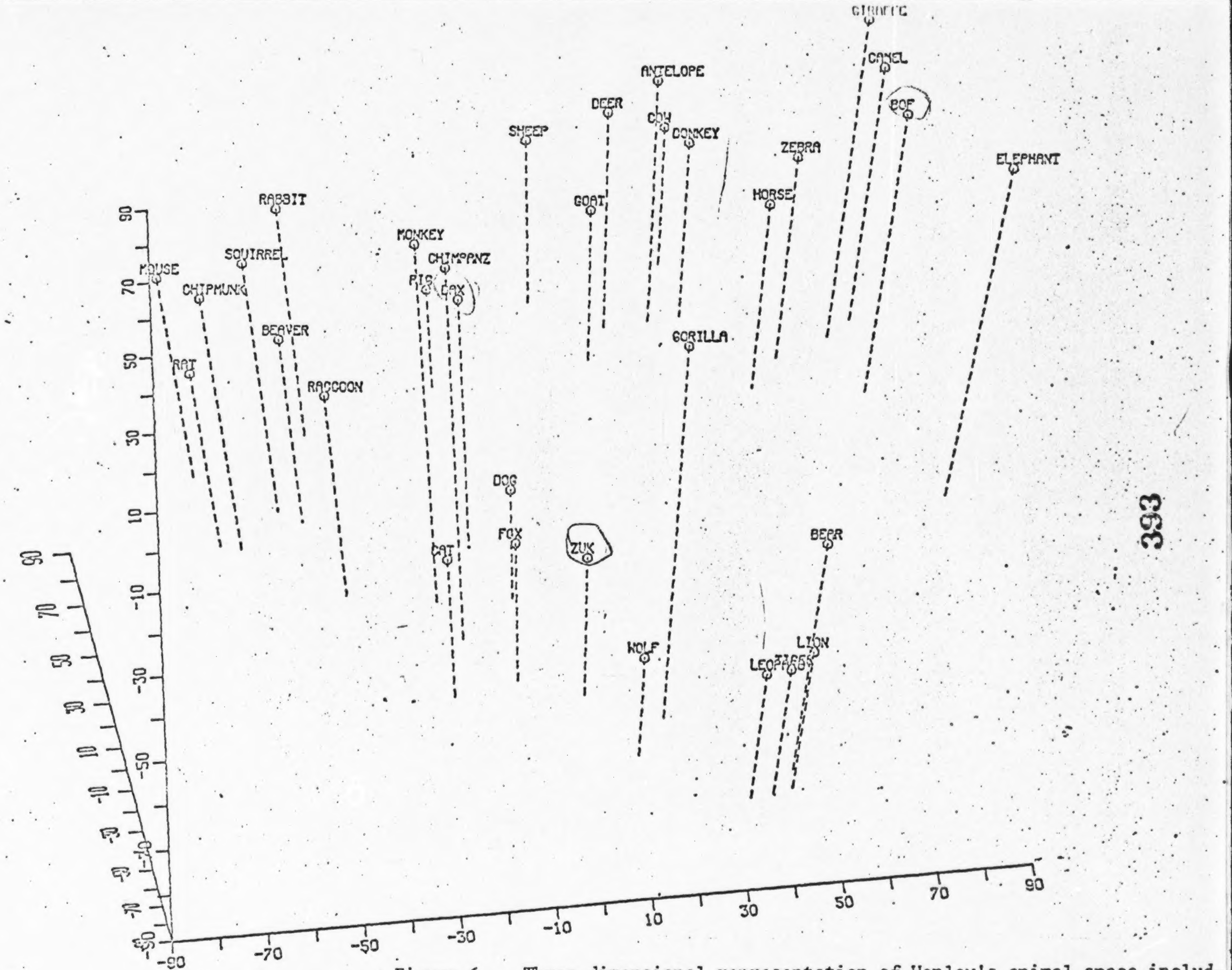


Figure 6. Three dimensional representation of Henley's animal space includ- the location of the three artificial animals, F, DAX and

Table 6

An anticipation training trial from Experiment III.

View one

1. SHEEP : CAT
BOF : ?

(a) Rabbit
(b) Chimpanzee
(c) Leopard
(d) Bear

View two

SHEEP : CAT
BOF : BEAR
BEST ANSWER: (d) Bear

Second (c) Leopard
Third (b) Chimpanzee
Fourth (a) Rabbit

study the subjects were given another analogy problem of the same form and the process was repeated. Table 6 shows an example of one trial of the analogy learning task. Response alternatives were chosen with the restrictions that the closest alternative was within .12 units of the ideal solution point; other alternatives included the most distant animal and two alternatives roughly Trisecting bisecting the distance from the ideal solution point to the most distant alternative.

Insert Table 6 about here

Twenty analogies were constructed to teach BOF, 18 to teach DAX and 17 to teach ZUK. Each of twenty-five subjects were given the training sequence on each of the three concepts. Following training, subjects were given rating sheets asking them to rate, on a scale from 1 to 10 (a score of 1 indicates that the animals are identical, a score of 10 that they are maximally different) the dissimilarity of each of the three artificial animals with each of the 30 animals in Henley's space and with each other. Following these ratings subjects were asked to give a verbal description of each of the three animals.

We thus have three measures of the degree to which the concepts which we have attempted to teach have in fact been learned by the subjects. Firstly, we can observe their ability to solve analogies while they are learning the concepts. Early in training we expect their responses will be as a chance level; later on, their responses should come to match those obtained in previous experiments. Secondly, we can compare our subjects' judgements of dissimilarity with those expected. Finally, we can look at their verbal descriptions to see the completeness of the concept and the extent to which different subjects describe the various artificial animals similarly.

Results and Discussion

Our first prediction regards the way in which these imaginary animals are used to solve analogy problems. If the subjects have really learned the concepts of BOF, DAX and ZUK their behavior during the latter phases of the training sequences should be indistinguishable from the behavior of our subjects in Experiments I and II. Since it appeared that learning was complete after, at most, five analogy anticipation trials, the anticipation data after the fifth trial should be the proper comparison. Table 7 shows the mean number of subjects choosing each of the four alternatives as the best. The average is taken over all anticipation trials after trial five on each of the three concept problems. The predicted values are those expected based on the parameter value of $\alpha = 1.68$ taken from Experiment II. The closeness of fit of these observed and expected values are definitely consistent with the idea that the concepts BOF, DAX and ZUK can be manipulated and used just as the other animal concepts used in our other experiments.

Insert Table 7 about here

In spite of the goodness of fit of these data these comparisons are indirect. It would seem that the similarity judgements would yield a more direct measure of the extent to which we have actually succeeded in teaching the concepts we set out to teach. Since BOF, DAX and ZUK are points exactly located in Henley's similarity space we can make direct predictions of dissimilarity ratings by measuring the distance between each artificial animal and each of the other animals in the space. We predict, in a strictly nonparametric way, that mean dissimilarity ratings will increase monotonically with distance in the space. Figures 7, 8 and 9 are plots of our observed dissimilarity rating

Table 7

Predicted and observed mean number of subjects choosing the ith closest alternative as the best. Predicted values generated under the assumption that $\alpha = 1.68$.

Rank of Alternative Distance From I							
1		2		3		4	
Pred.	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.	Obs.
12.4	13.3	6.6	8.2	3.1	2.5	1.6	1.6

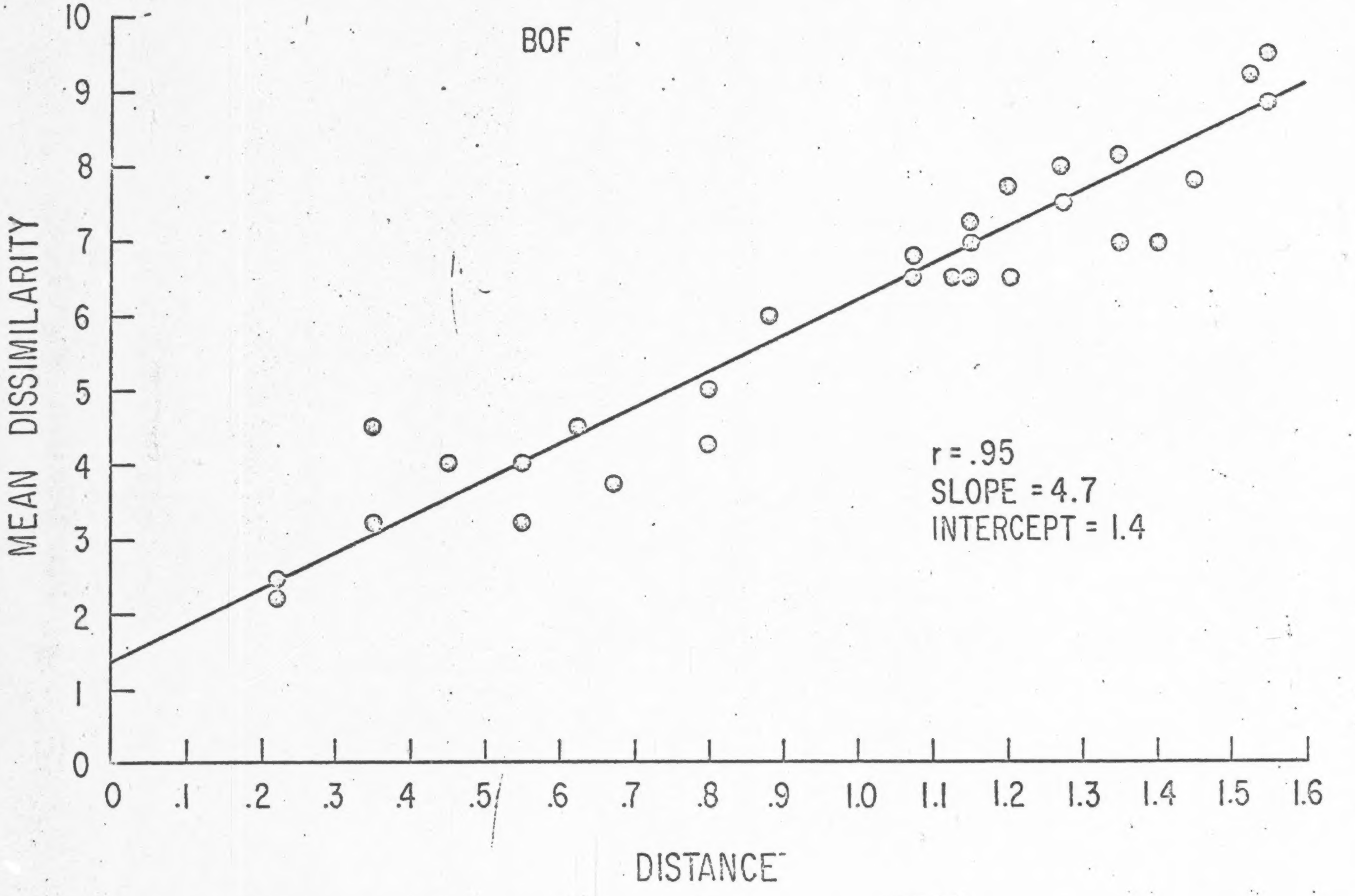


Figure 7. Mean dissimilarity judgments between BOF and each of the thirty mammals as afunction of the distance of the mammal from BOF.

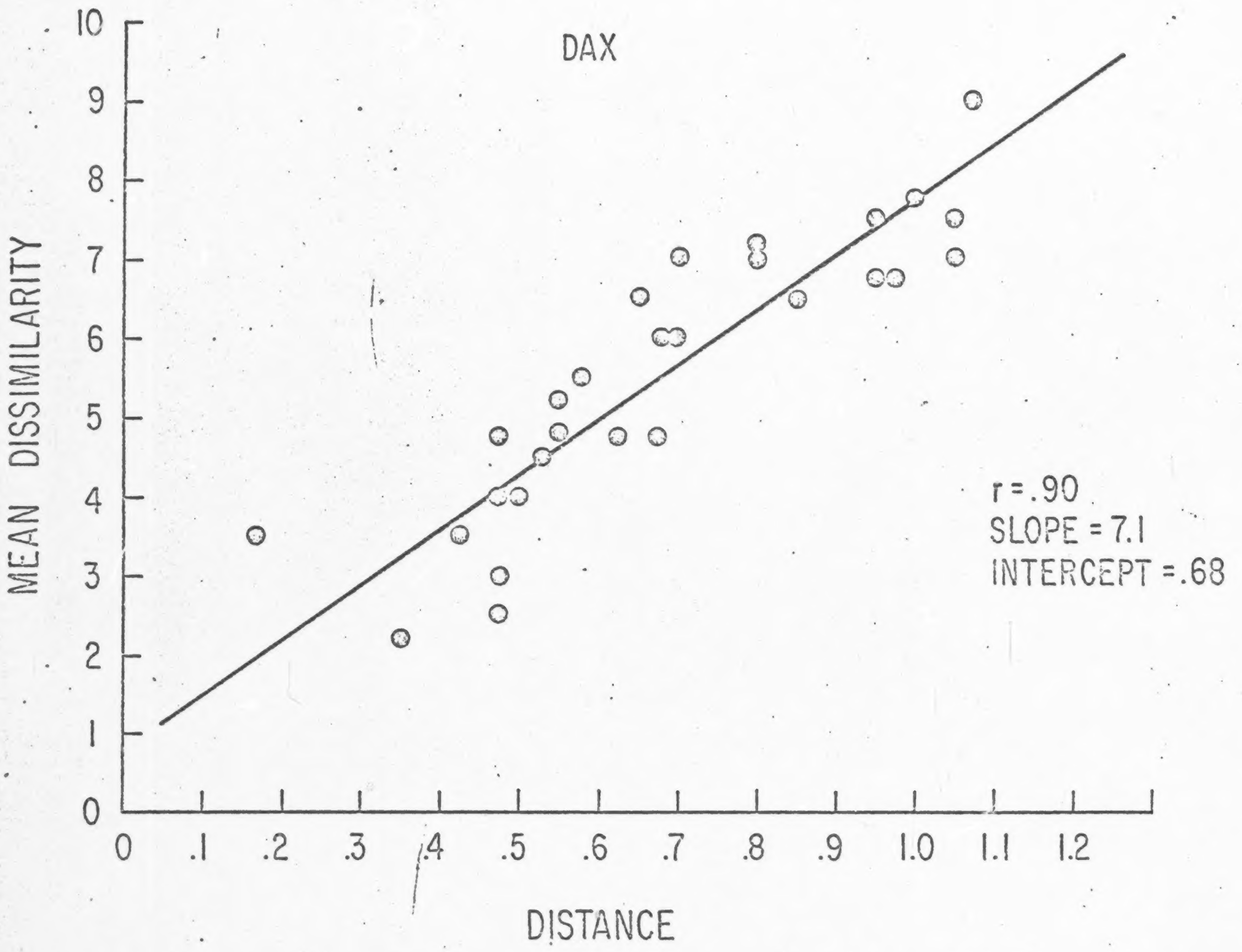


Figure 8. Mean dissimilarity judgments between DAX and each of the thirty mammals as a function of the distance of the mammal from DAX.

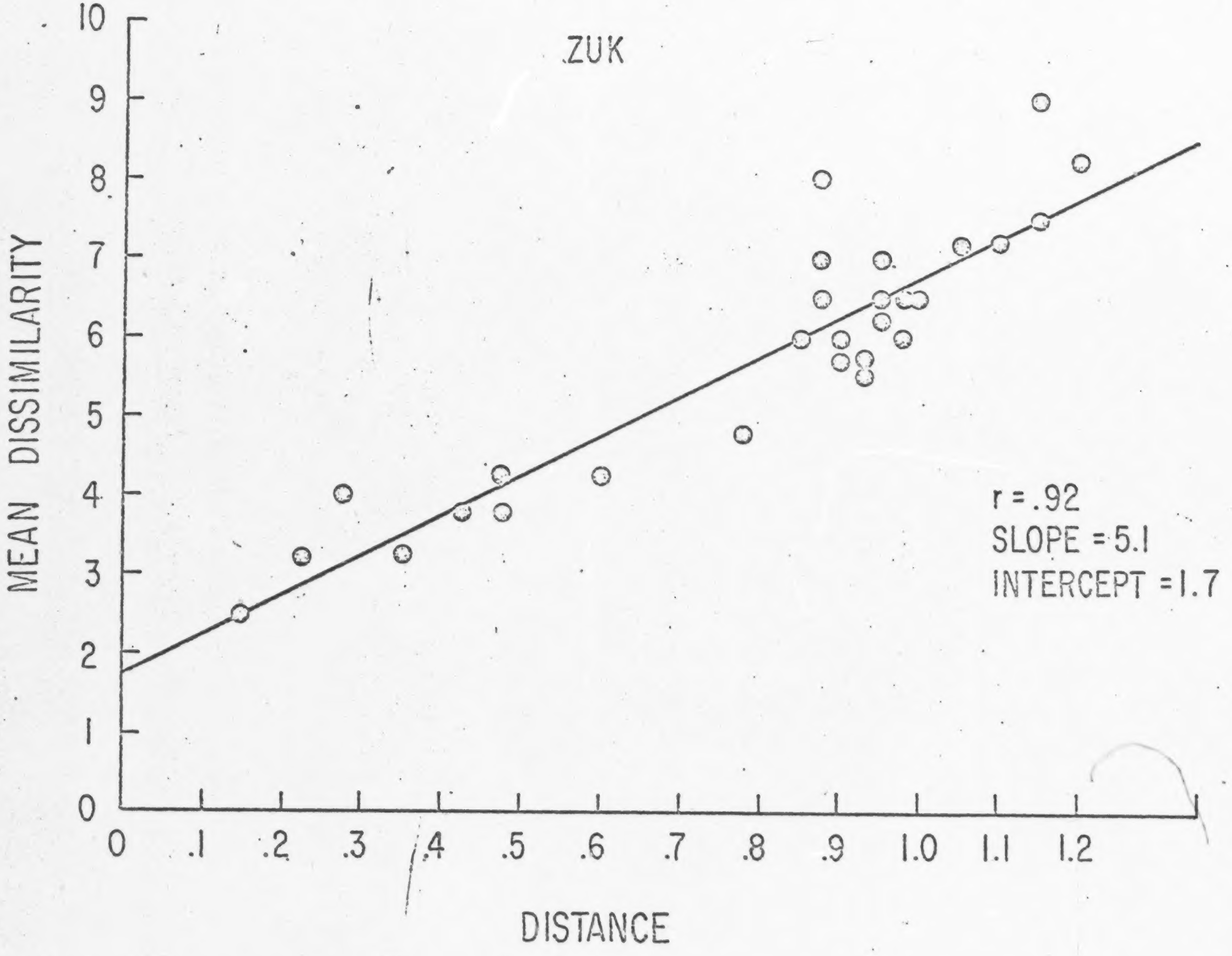


Figure 9. Mean dissimilarity judgments between ZUK and each of the thirty mammals as a function of the distance of the mammal from ZUK.

Insert Figures 7, 8 and 9 about here

versus distance measured in Henley's space. The product moment correlations were r = .95, .90 and .92 for BOF, DAX and ZUK, respectively. The Kendal's tau values were T = .79, .75 and .73 for the three animals.

Although on an a priori basis correlations between .9 and .95 seem good it is not clear how good the correlation should be in order to be comparable to real concepts. To help answer this question, Nancy Henley kindly supplied us with her raw data for comparison. We were thus able to compute similar correlation coefficients between her subjects' dissimilarity judgements and her derived semantic space. These coefficients were found to vary between r = .98 for camel, chipmunk and squirrel, and r = .86 for gorilla. The values of tau varied between T = .91 for mouse and T = .58 for bear. Furthermore, of the thirty animals, 11 had values of r higher than .95 and 6 had values lower than .90. The remaining 13 had values between .90 and .95. The conclusion thus seems clear. The dissimilarity judgements given by our subjects comparing BOF, DAX and ZUK to each other and to each of the other thirty animals are every bit as predictable as are subjects' responses with real animal names.

One final computation of interest was carried out with the dissimilarity data. We added our subjects' data to that of Henley, thus increasing the number of animals from 30 to 33. We then inserted the data into a version of TORSCA[8A] to obtain a revised scaling. Figures 10 and 11 show a superposition

Insert Figures 10 and 11 about here

Figure 10. A superposition of dimensions 1 and 2 of Henley's original scaling and the scaling generated from including data from Experiment III, shows indicated change from original to revised scaling.

FILMED FROM BEST AVAILABLE COPY

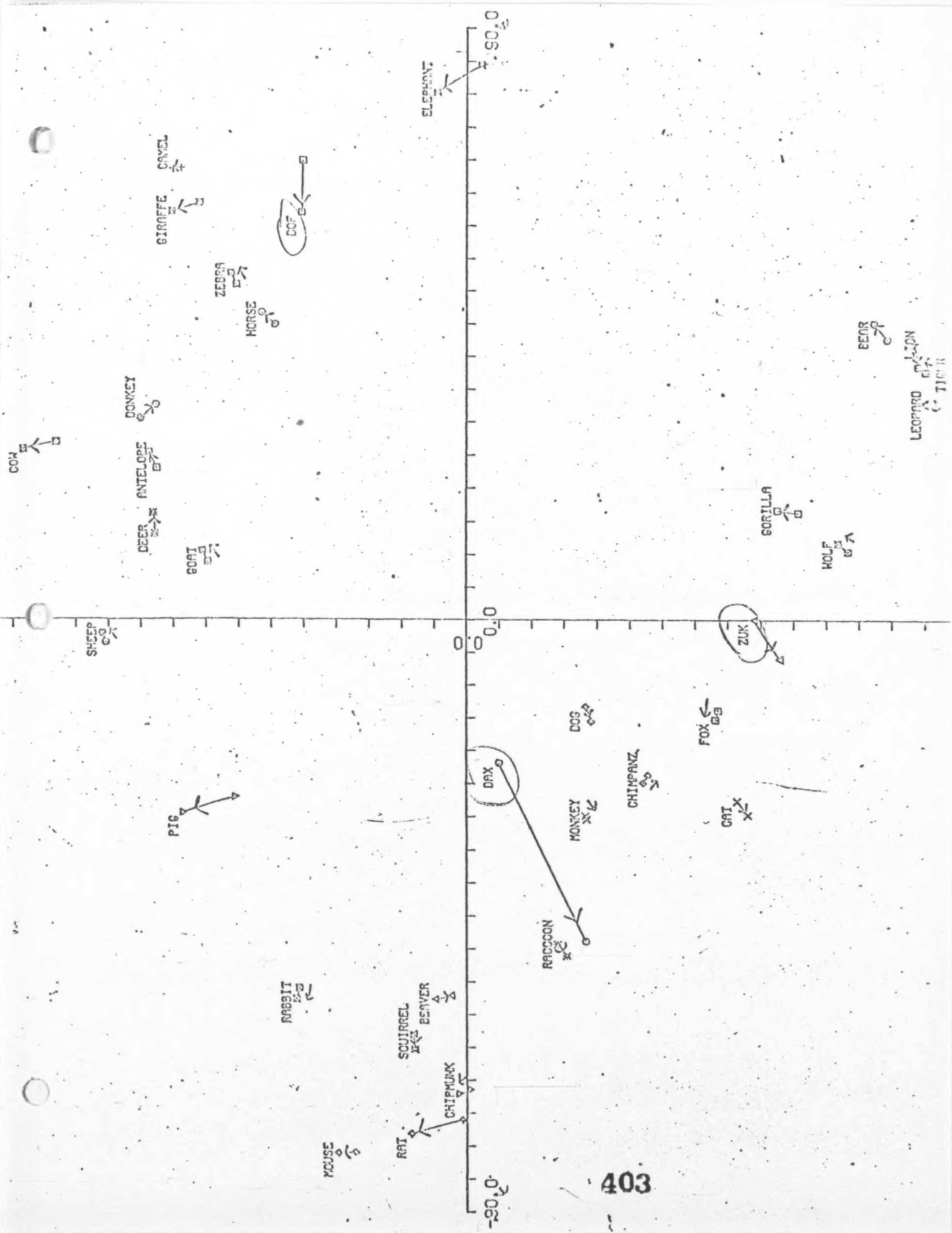
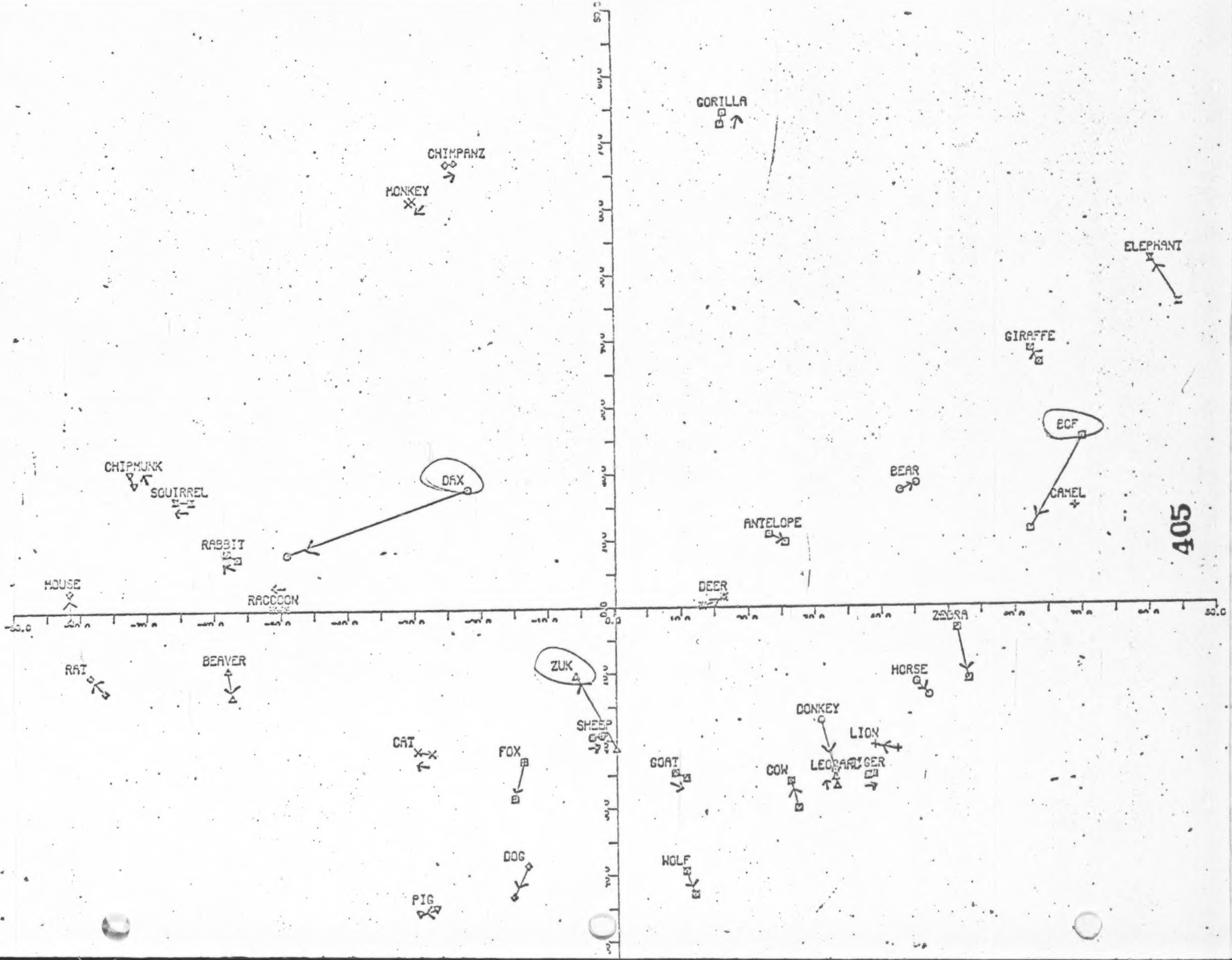


Figure 11. A superposition of dimensions 1 and 3 of Henley's original scaling and the scaling generated from including data from Experiment III, shows indicated change from original to revised scaling.


GORILLA
CHIMPANZ
MONKEY
ELEPHANT
GIRAFFE
ECF
BEAR
CAMEL
CHIPMUNK
SQUIRREL
DAX
ANTELOPE
RABBIT
MOUSE
RACCOON
DEER
ZEBRA
RAT
BEAVER
ZUK
HORSE
DONKEY
SHEEP
LION
CAT
FOX
GOAT
COW
LEOPARD
TIGER
DOG
WOLF
PIG
-50.0
0.0
10.0
50.0

of the revised scaling on Henley's original semantic space. The arrows show the migration of each of the points from the first scaling to the second. A value of Kruskal's of stress index of 10% was obtained in the rescaling as compared to a value of 9.4% in Henley's original scaling. Of special interest is the movement of the three artificial animals. The origin of the arrow indicates the intended location of the concept; the end of the arrows indicates the final location of the concepts as indicated by the rescaling. BOF moved a total of .153 units, DAX moved .277 units and ZUK moved 13.2 units. The largest movement of the animal concepts occurred with elephant which moved .094 units. The fact that the artificial animals moved more than the real animals should not be surprising on three counts. Firstly, and most obviously, the vast majority of the data going into the rescaling was the same data which determined the location of the animals in the first place. Secondly, it should be recalled, the training analogies were correct with a tolerance of .12 units. Hence, our training procedure insures that we will teach what we want only within .12 units. Thirdly, although the scaling procedure requires that the same monotonic transformation be applied to all of the dissimilarity judgments in a given scaling. Our procedure tended to promote context-effects. That is, our subjects were given a sheet asking them to judge the dissimilarities of DAX and antelope, DAX and bear, etc., down to DAX and zebra. Then a similar procedure was used for the BOF judgments, etc. It is thus quite likely that context-effects will emerge. A glance at Figures 7, 8 and 9 indicate the problem. Although the regression line fits very well in all three cases, the slope for DAX is about 50 percent higher than for BOF and ZUK. The reason seems to be that DAX is a relatively central animal and thus there are no animals very distant from DAX. Nonetheless, the entire range 1 to 10 is used. Hence, the higher slope.

The final datum from Experiment III is the verbal descriptions. It is never clear what should be made from such data; nonetheless, they occasionally make interesting reading. To that end, Tables 8, 9 and 10 give a compilation of subjects' comments regarding their ideas of the three animals.

Insert Tables 8, 9 and 10 about here

General Discussion and Conclusions

We began our discussion with an analysis of reasoning. We suggested that reasoning was the collection of processes or algorithms which operate on organized memory in various information retrieval tasks. Any particular reasoning task, such as analogical reasoning, is the application of a particular set of algorithms to perform a particular type of data retrieval. The problem of characterizing a reasoning process is simply the specification of a data base to retrieve a particular kind of information. We chose to characterize the data base for our theory of analogical reasoning as a multidimensional Euclidean space. Given that specification of the data base our choice for a reasoning algorithm becomes very natural. It would appear that the results of Experiments I, II and III confirm the accuracy of our description of analogical reasoning among animal concepts. There are many possible ways to generalize the results we have found. The most straightforward of these would be to search for other sets of ~~elements~~ concepts which are well embedded in a multidimensional similarity space and show that our theory also gives an accurate description for analogy formation among these concepts ~~behavior-in-that-context~~. We have, in fact, tried some other spaces. In particular, we have generated analogies from the Munsell color space and have

VERBAL DESCRIPTION: BOF (18 subjects)

SIZE:
- 5 Large
- 5 Camel-sized
- 3 Horse-sized
- 1 Rather large
- 1 Relatively large
- 1 Fairly big
- 1 Medium-sized mammal
- 1 Bigger than donkey
- 1 Very large antelope
- 1 Somewhat smaller than elephant
- 1 Larger than elephant

SHAPE:
- 1 Slender
- 1 Heavy
- 1 Visualize as big donkey

APPENDAGES:
- 1 Antlers or horns
- 1 Probably has horns
- 1 Small horns
- 1 Horns or humps

LEGS:
- 5 Four legs
- 1 Uses all four legs
- 2 Long legs
- 2 Hooved feet

NECK:
- 2 Long
- 1 Extended but shorter than giraffe
- 1 Shorter than giraffe
- 2 Hooved feet

HAIR:
- 2 Short
- 2 Fur
- 1 Rather long (unlike camel, giraffe, cow)

WILDNESS:
- 2 Wild
- 1 Completely domesticated
- 1 Could be used as work animal

MOVEMENT:
- 1 Swift moving
- 1 Runs very fast (like antelope)
- 1 Much more agile than camel
- 1 Not as coordinated as giraffe
- 1 Doesn't move very swiftly
- 1 Strong (like elephant or camel)

LIVE:
- 1 In desert or jungle
- 1 On plains

FOOD:
- 1 Non-carnivor
- 1 Probably eats plants
- 1 Grazes
- 1 Naturally eats grain and leaves

T :
- 1 Mammal

INTELLIGENCE:
- 1 Not as intelligent as giraffe

LIFE STYLE:
- 1 Travels in herds with leader
- 1 Fairly mobile (like giraffe or zebra)

SIMILAR TO:
- 6 Camel
  - 1 Features similar to camel
  - 1 Camel
  - 1 Cross between camel and horse
  - 1 Somewhere between giraffe and camel
  - 1 Similar to camel or giraffe, closer to giraffe
  - 1 Cross between camel, giraffe, zebra, elephant
- 5 Giraffe
  - 1 Like giraffe but shorter neck (mentioned camel first)
  - 1 Somewhere between giraffe and camel
  - 1 Similar to camel or giraffe, closer to giraffe
  - 1 Very similar to domesticated giraffe
  - 1 Cross between camel, giraffe, zebra, elephant
- 2 Goat
  - 1 Cross between goat and deer but camel-sized
  - 1 Something like goat or sheep
- 2 Antelope-Deer
  - 1 Very large antelope
  - 1 Cross between goat and deer but camel-sized
- 1 Cross between camel and horse
- 1 Visualize as big donkey
- 1 Something like goat or sheep
- 1 Cross between camel, giraffe, zebra, elephant
- 2 Gorilla
  - 1 Characteristics similar to gorilla
  - 1 Possibly slight similarities to a primate

Table 9

VERBAL DESCRIPTION: DAX (18 subjects)

SIZE:
- 6 Small
- 2 Medium-sized
- 1 Rather small
- 1 Very large
- 4 Size of beaver
- 1 Larger than beaver
- 2 Size of rabbit
- 1 Somewhat larger than rabbit
- 1 Size of squirrel
- 1 Size of raccoon
- 1 2-feet long
- 1 20-30 pounds

LEGS:
- 2 Four legs
- 3 Can stand on two legs
  - 1 Can support self on two as well as four
  - 1 Can manipulate with front feet while standing on rear
  - 1 Can stand on hind legs
- 1 Short legs
- 1 Has claws

APPENDAGES:
- 1 Long bushy tail
- 1 Long husky tail
- 1 Short tail

HAIR:
- 4 Furry
- 1 Long fur
- 2 Hairy
  - 1 Like chimpanzee

FEATURES:
- 1 Small ears, bright eyes, all that
- 1 Large front teeth
- 1 Big ears
- 1 Features of rabbit
- 1 Obviously green

WILDNESS:
- 1 Possibly domesticated
- 1 Close to being domestic; lives around man (like mouse)
- 1 Can be seen in wild or kept as pet (like rabbit)

MOVEMENT:
- 1 Fairly quick
- 1 Fast-moving
- 1 Quick-moving, agile
- 1 Can climb trees well (like raccoon)

LIVE:
- 2 In forest (perhaps trees); doesn't like water as does beaver
- 2 Near water
  - 1 Often near water
  - 1 Dexterous in water, like beaver

FOOD:
- 1 Herbivorous or possibly omniverous
- 1 Eats young plant shoots

TYPE:
- 5 Rodent family
- 1 Mammal

INTELLIGENCE:
- 1 Clever
- 1 Dumb

PERSONAL CHARACTERISTICS:
- 1 Industrious
- 1 Cuddly
- 1 Fairly timid, not agressive

SIMILAR TO:
- 4 Rabbit
  - 2 Larger than rabbit, however
  - 1 Like rabbit in life and size
  - 1 Like rabbit
- 2 Beaver
  - 1 Larger than beaver, however
  - 1 Cross between beaver, chimp, raccoon
- 2 Raccoon
  - 1 Very similar to squirrel and resembles raccoon
  - 1 Cross between beaver, chimp, raccoon
- 1 Squirrel
  - 1 Very similar to squirrel and resembles raccoon
- 1 Chimpanzee
  - 1 Cross between beaver, chimp, raccoon

Table 10

VERBAL DESCRIPTION: ZUK (19 subjects)

SIZE:
- 3 Small
- 3 Medium
- 1 Medium small
- 1 Largish
- 2 Size of fox
- 1 Same or smaller than fox
- 1 Size of large fox; smaller than donkey
- 1 Larger than fox; smaller than lion
- 1 Size of lion
- 1 Size of dog
- 1 Larger than dog
- 1 Slightly smaller than wolf
- 1 Size of raccoon or beaver, like DAX

SHAPE:
- 1 Similar in proportion to dog
- 1 More like dog or cat than zebra in build and fur; maybe like a great big mouse

HAIR:
- 2 Furry
- 1 Moderately hairy
- 1 Probably has shaggy fur (longer than fox)
- 1 Long-haired coat like goat and wolf
- 1 More like dog or cat than zebra in fur

APPENDAGES:
- 1 Tail
- 1 No horns; long flat tail

LEGS:
- 1 Four legs
- 1 Four stubby legs

FEATURES:
- 1 Similar to dog in features
- 1 Probably has unusual colorings
- 1 Large front teeth

WILDNESS:
- 2 Wild
- 1 Fierce
- 1 Attacks smaller animals
- 1 Clearly not domesticated; less wild than fox
- 1 Fairly aggressive, or at least mischievous and prying

MOVEMENT:
- 1 Quick
- 1 Fast runner
- 1 Very fleet, swift-moving
- 1 Not as fleet as fox
- 1 Slow
- 1 Agile like wolf

LIVE:
- 1 Probably in forest
- 1 In woods; in ground or caves or hollowed trees; uses left-over tunnels
- 1 Perhaps in large holes in ground
- 1 Excellent on mountains, like goat

LIFE STYLE:
- 1 Not in packs like wolf; good family life

FOOD:
- 4 Predatory
    - 1 Seeks prey (like cat)
    - 1 Attacks smaller animals
    - 1 Preys on chickens and sometimes sheep, goats; problem to farmer
- 2 Carniverous

TYPE:
- 1 Great cat family
- 1 Probably canine; possibly feline
- 1 Dog family
- 1 Mammal

INTELLIGENCE:
- 1 Clever
- 1 Not as intelligent as fox

SIMILAR TO:
- 5 Cat family
    - 1 Cat-like (size largish)
    - 1 More like dog or cat than zebra in build and fur
    - 1 Cross between cat and canine families
    - 1 Like lion or other large cat
    - 1 Maybe is cheetah
- 4 Canine family
    - 1 Like dog
    - 1 More like dog or cat than zebra
    - 1 Cross between fox and dog
    - 1 Cross between cat and canine families
    - 1 Like wild dog
    - 1 Cross between goat and wolf
- 3 Fox
    - 1 Resembles fox
    - 1 Cross between fox and dog
    - 1 Most like fox or any wild dog
- 2 Farm animals
    - 1 Cross between goat and wolf
    - 1 Horse-like
- 1 Maybe like a great big mouse

found essentially similar results.[6] Problems arise when relations among words are not well represented in a multidimensional space. (Hierarchical relations are examples.) These kinds of examples seem to demand a more general representation ~~in~~ of memory structure which is capable of handling more complex sorts of relationships while still holding the multidimensional representation as a sub-case. It is interesting in this regard that Quillan's (1969) teachable language comprehender (TLC) seems to be such a generalization. If Quillan encoded a set of words, all with the same superset and which differ~~ing~~ from one another only with regard to values on a common set of attributes, the ~~information~~ words could be equivalently encoded into a multidimensional representation. It is perhaps not coincidental that the set of words Henly scaled ~~fits~~ have exactly these ~~requirements~~ characteristics. We are presently engaged in finding ways of specifying analogical relationships within a data structure of this more general nature.

Footnotes

1. The research reported in this paper was supported in part by GB-8587 from the National Science Foundation and by Grant NG-07454-03 from the National Institute of Health, the United States Public Health Service.
2. Experiment I was carried out by Sharon Wilson as part of an undergraduate research project under the sponsorship of the first author.
3. In cases in which the expected value was less than 5 cells were collapsed. The X^2 value was computed over this reduced matrix. Hence different comparisons have different degrees of freedom.
4. Thanks are due to Richard Meltzer who provided a copy of TORSCA and was otherwise instrumental in carrying out this analysis.
5. This experiment was carried out by Glenn Rice as part of a research project under the sponsorship of the first author.

Semantic Structure: Psychological Evidence for Hierarchical Features[1]

Kenneth N. Wexler
University of California, Irvine

This paper was prepared for the Research Workshop on Cognitive Organization and Psychological Processes, sponsored by the Council on Basic Research in Education, held in Huntington Beach, California, August, 1970.

Abstract

Using a set of "have" verbs as an example, the hypothesis that semantic structure is hierarchical is tested. The experiments are triad tests in which a subject is asked to pick, out of 3 words presented, the one most different in meaning from the other 2. Hierarchical structure and process models are developed to test the results. Experiment I is a triad test on 9 verbs. Experiment II extends the method to sentences. Experiment III measures response times to the word triads and thus provides evidence that a hierarchical structure and process model is correct. Evidence is presented to show that a "lexicon" model rather than a "tree" model is the appropriate psychological representation of semantic structure.

Over the past year I have been working in a field to which I keep giving different names. Sometimes I call it "semantic memory". Other times it's the "structure of the lexicon", or "semantic structure" or "hierarchical structures and processes". A goal of this paper is to show in what way these terms are equivalent and to tease apart differences where they exist.

Essentially the question I am asking is, what is the cognitive (subjective) organization of the meaning of words? The relevance of this question to, say, the study of memory is obvious. If one wants to be shown this experimentally he need only look at the work of Bower (1970) or Kintsch (1970). But we still have no clear idea of how the meaning of words is organized. I think that without a theory of the structure of the lexicon work on memory processes will be hampered. For if one is trying to investigate memory processes which are not well understood, and these processes apply in some way to semantic structures which are not understood, then there is little chance of achieving understanding. In fact it was exactly these considerations that led psychologists to try to eliminate meaning from their experiments by substituting nonsense syllables for words. However, recently there has been much more research using words and other "natural" materials. A complete explanation of why this is so would require a history I am not prepared to carry out. However, I suspect that there are at least two reasons. First, many psychologists may have come to believe that many (if not most) memory processes lose their most interesting properties when meaning is eliminated from the materials upon which they work. Second, work on materials with meaning was probably stimulated by the challenge of linguists and others to explain psychological processes on "natural" materials, that is, materials with which human beings actually deal outside the laboratory.

These considerations suggest that a direct attack on semantic structure, bypassing problems of memory, would be a fruitful approach to take, even if one is primarily interested in memory. Thus, although I do not report any experiments of the classic type on memory in this paper, I hope to convince the reader that the results are relevant to the study of semantic memory.

The experimental tool with which I have been working is the "triad test". A subject is given 3 words and told to select which one is "most different in meaning from the other two". This experiment was first introduced with the "meaning" instructions by Romney and D'Andrade (1964). This experiment certainly would not usually be considered a memory experiment. Rather it would probably go under the rubric of an experiment in "judgment of similarity". Nevertheless, the reader will surely note that long-term memory is intimately involved. The data show that the subject makes his choice based on the meanings of the words. This meaning is not presented to him directly (unless one believes rather strongly in phonetic symbolism). Rather, he must search his memory for the meaning, given only the word. Also, the subject must make his choice of the most different word based on the meaning he has retrieved. It is an open question, of course, whether the process by which the subject does this is similar in any way to the process by which he retrieves and processes meanings in other kinds of experiments or, indeed, in real-world, natural language processing.

The first experiment I did using the triad test was a replication of the Romney and D'Andrade kinship experiment. The words were the 8 basic English male kinship terms (father, brother, etc.). Let me quickly summarize the results. See Wexler and Romney (1970) for the details. We distinguish between "structure" and "process". The "structure" of the semantic domain is a paradigmatic or grid model. That is, the structure is exactly like a "concept" in psychology. There are dimensions and values. One dimension might be "generation", with values 0,1,2. The "process" is a model of what a subject does in a triad experiment given a certain structural model. The process model says calculate the city-block (grid) distance between each pair of words in the triad (there are 3 pairs) and classify together the pair with smallest distance, that is, pick as most different the other (third) word. If there is a tie for smallest distance, guess randomly.

This process model was applied to 2 different structural models and one of the models clearly did a better job of predicting the data than the other. In fact, the better model fit the data fairly well, especially since there were no parameters. Thus, for the kinship case, paradigmatic or multi-dimensional models of the structure seemed to be appropriate.

I was left uneasy about the generality of these results for the study of the semantic structure of ordinary words because, intuitively, kinship seemed to be a special case (though it is not easy to give reasons why this is so). In fact, anthropologists, with some exceptions (e.g., Berlin,), have mostly rigorously analyzed only kinship, pronouns and color terms. Thus I was excited when I came upon the work of Bendix (1966), who gave a paradigmatic analysis of a set of "have" or "transfer" words, like "get", "give", "borrow". Experiment I is a triad experiment on these words. This experiment was reported and analyzed elsewhere (Wexler, 1970, mimeographed), so I do not want to give all the details of the analysis, but it is necessary to present it in order to understand what I am up to in the rest of this paper.

Experiment I

Subjects. There were 24 subjects, undergraduates at the University of California, Irvine, who were required to participate in experiments. Subjects were run in groups of 1 to 3 subjects per group.

Words. All the items studied by Bendix were used, with the exception of "get rid of", because this item was composed of more than one word. There were 9 words used. They were "get", "find", "give", "lend", "borrow", "take", "lose", "keep" and "have".

Procedure. The 9 "have" verbs were presented to subjects in triads, that is, groups of 3. All $\binom{9}{3} = 84$ triads were used. A subject was told to select the one word of the three in each triad which was "most different in meaning from the other two". The triads were presented visually, on sheets of paper 24 triads to a sheet (12 on the last sheet). The triads were arranged in random order, the same order for each subject. Subjects were self-paced, going on to the next triad as soon as they had finished the preceding one. The task took from 20 to 30 minutes to complete.

Results. First let me define a simple statistic. Suppose in a triad XYZ the term X is selected as "most different". Then we say that Y and Z are "classed together". A convenient statistic to consider is C, the number of times a subject classes together a given pair over the entire list of triads. Since there are 9 words in the experiment, a given pair appears in 7 triads, that is, each of the other 7 words appears once with the pair to make a triad. Thus values of C can range from 0 to 7. For each pair, we calculate E(C), the mean value of C over subjects. The calculation is done simply by finding the probability of classing together the pair in each of the 7 triads in which the pair appears and summing the results. In other words, $E(C) = \sum_{i=1}^{7} p_i$, where p_i is the probability of classing together the pair in triad i, for i ranging over the 7 triads in which the pairs appears. This is easily shown to yield the correct result. The proportion of subjects classing together a pair in a triad is taken as an estimate of the probability of classing together the pair in that triad.

C may by thought of as a measure of similarity of 2 words, and while it does not tell the whole story, it captures a good deal of what is happening in the experiment. The values of C obtained in Experiment I are listed in Table 1.

Insert Table 1 here

Let us consider the following simple model for part of the semantic structure of the verbs. This paradigmatic model is diagrammed in Table 2.

Insert Table 2 here

The model is inspired by Bendix' analysis, but is quite different from it in detail. What is carried over from Bendix is the notion of considering an instant t of time as a reference point and letting features refer to the state of the system before and after t. Thus, the first feature refers to the state of possession before time t and has values "Have" and "Not Have". The second feature refers to the state of possession after time t, and we label the features 'Have" and "Not Have", just as we did for the first feature. In order to get the intuitive feel for this analysis, it is best to embed the verbs in a sentence context, for example, "Andy verb the book (preposition) Charley". Thus, in the sentence "Andy gives the book to Charley", there is a time t such that before t Andy has the book and after t Andy doesn't have the book. Thus the first feature of "give" has the value "Have"(H) and the second feature has the value "Not Have"($\bar{H}$). We can write this easily in the form: give=$H\bar{H}$. In the sentence, "Andy borrows the book from Charley", there is a time t such that before t Andy doesn't have the ball and after t Andy has the ball. Thus, borrow=$\bar{H}H$. In the sentence, "Andy has the book", no transfer of property takes place, but we can imagine an instant of time t and apply the features before and after t. Thus, have=HH. Bendix did not use any verbs which fit into the $\bar{H}\bar{H}$ quadrant, but such verbs exist, for example, "want" and "need".

We can now apply the grid process model to the paradigmatic structural model. A few sample caluclations will be given. Let D be the city-block distance. Consider the triad (have,get,borrow). D(have,get)=1, D(have,borrow)=1, D(get,borrow)=0. Thus this last distance is minimal, and "get" and "borrow" are classed together, that is, the model predicts that "have" is chosen as most different with probability 1. For (borrow,keep,give), we have D(borrow,keep)=1, D(borrow,give)=2, D(keep,give)=1. Thus (borrow,keep) are each classed together with probability 1/2, that is, "give" and "borrow" are each chosen as most different with probability 1/2.

Once we have made the predictions for triads from this model, we can calculate the theoretical values of C. These figures are given in

Table 1. Overall, the fit between theory and data is not bad. The Pearson product-moment correlation between theory and data is r=0.90.

The paradigmatic model is crude in the sense that it doesn't distinguish among all the verbs. Thus, for example, "take", "get", "borrow", and "find" all have the same values, and there is no way in this model to distinguish them from one another.

It is intuitively clear, however, that these words are different. The problem now arises of how to differentiate among the verbs formally. We could continue in the same way as before and extend the paradigmatic, multi-dimensional model by adding more features of meaning. There are two immediate problems with this suggestion. First, the features involving possession which we have already discussed seem intuitively more important than any others and should be marked in some special way. In fact, we have seen that these features alone do a rather good job in predicting our experimental results. Second, it is difficult to think of plausible features (besides those already used) which would apply to all the verbs and would discriminate among them. If we tried to list plausible features, we would probably have to think of features on which a few of the verbs were "marked" and then give the verbs either a marked or unmarked value on those features.

In order to solve this problem, I suggested a new kind of model, an "embedding structure" (Wexler, 1970). An embedding structure is a generalization of a tree or hierarchy. Essentially it is a hierarchy with cross-classification possible at each level or node. Thus the paradigmatic model above might be the first "cut" or level of structure and then each of the sub-groups (e.g., "give", "lend", "lose") would be analyzed further.

How would a process model work on such a structure? There is no defined city-block distance over all pairs, so we need another model. A natural model does exist. We won't need the general concept of embedding structures, so to simplify matters we assume that the structure is simply a binary tree, which is a special case of embedding structures. The terminal nodes of the tree are the words, and the other nodes are "features". The hierarchical process model works down from the root.[2] It starts by looking at the first level of features. If the 3 words are different on this cut, then a decision is made here. It must be, for a binary tree that, if there is a difference then 2 of the words are the same and 1 is different. The one that is different is chosen as most different. Suppose, however, that the 3 words are all the same at the first level. Then the model goes to the next level and repeats the process. In other words, the model looks for the first difference and makes the decision, ignoring what comes later.

I want to emphasize that both the structural and process models have been defined formally in Wexler (1970), not only for binary trees, but for trees in general, and, more generally, for embedding structures. Let me motivate an example of a tree structure. If one looks at the values of C in Table 1, he will notice that pairs with either "have" or "keep" as one of their members, and "give", "lose", or "lend" as the other, tend to be over-predicted. For example, C(give,have)=0.58 in the data, while the theoretical value is C=2.00. One way to capture this phenomenon would be to weight the dimensions in the paradigmatic model. Another way to do it, possibly without estimating a parameter (depending on your point of view) is to construct the tree model shown in Figure 1. The structures of the words may then be

---

Insert Figure 1 here

---

defined in the following way, as in Wexler (1970). Give the value 0 to a left branch and 1 to a right ranch. Then we have

have=keep=(0),(0),

get=take=borrow=find=(0),(1),

give=lend=lose=(1).

These equations define a semantic model, which we call the "hierarchical model".

A few sample calculations: For (have, get, give), there is a difference already at the first level, since "give" has value 0, while the others have value 1. Thus "give" is most different. For (keep, take, borrow), all 3 words have value 0 at the first level, thus the model goes to the second level. Here "keep" has value 0 while the other 2 words have value 1. Thus "keep" is chosen as most different.

The predictions for C for the hierarchical model are shown in Table 1. The Pearson r between theory and data is r=0.97. When all 3 words in a triad were from the same group, (e.g., get, take, borrow) the model assigns probabilities of 1/3 to each word. In my previous paper I went on to show how the model could be completed, that is how the structure could be extended so that all the words could be distinguished. This involved a consideration of the actual triad choices, not just the C values. It also pointed up the difficulty of constructing a multi-dimensional, paradigmatic model to explain the data. I won't repeat this analysis here since it should be clear how a complete hierarchical analysis can be given.

What we have then is a semantic analysis which is a hierarchy. Alternatively, we have a lexicon, with each word given a sequence of features. Later I will make some comments about the extent to which these analyses are not equivalent.

It thus looks as if Bendix' paradigmatic analysis is not correct. However the paradigmatic analysis that we gave is not exactly Bendix' analysis. The problem is that Bendix analyzed the words in sentence contexts, that is, frames, and we have analyzed the words in isolation. Experiment II (done jointly with Gregory Truax) was undertaken to study the items exactly as Bendix presented them and also to see if once again hierarchical structure prevailed. Split-half reliability studies were also made.

Experiment II

Subjects. Subjects were from 2 undergraduate Anthropology classes at San Fernando Valley State College. The classes had not studied any of the materials or methods used in the experiment. Originally there were 75 subjects. Two of the subjects (one from each class) misinterpreted the task, and their data was discarded. Three other subjects failed to complete the task. Their data was also discarded. Thus 70 subjects remained, 30 in class 1 and 40 in class 2. The experiment was run on all the members of a class at the same time. Both classes were run on the same day.

Sentences. There were 11 sentences, obtained by substituting a constant word for the variable in each of Bendix' frames. The substitutes were "Andy" for A, "the book" for B, and "someone" for C. "Someone" was chosen instead of a definite noun phrase because many of the sentences did not contain C, and we supposed that "someone" would add less semantic content than, say "Charley", or "the man". The sentences used as items in the triad test are listed on the left below, with the Bendix forms from which they are derived on the right.

Andy gets the book.	A gets B
Andy finds the book.	A finds B
Someone gives Andy the book.	C gives A B
Someone gets Andy the book.	C gets A B
Someone lends Andy the book.	C lends A B
Andy borrows the book from someone.	A borrows B from C
Andy takes the book from someone.	A takes B from C
Andy gets rid of the book.	A gets rid of B
Andy loses the book.	A loses B
Andy keeps the book.	A keeps B
Andy has the book.	A has B

Exactly why Bendix chose these forms is not clear from his work. For example, why is A sometimes the subject of the sentence (A takes B from C) and sometimes the indirect object (C gives A B)? The best guess I can make is that Bendix wanted to mark as many forms as possible positive on the feature "A has B after time t".

Procedure. All $\binom{11}{3}$=165 triads were presented to each subject. The presentation was the same as in Experiment I, on sheets of paper. The subject was told to select the sentence most different in meaning from the other two. As before, no clarification was given if questions were asked about the instructions.

The left to right order of the sentences was randomized within triads so that a given sentence appeared in the first, second or third position approximately the same number of times in the 165 triads. Two randomizations of the order of the triads were used. This was to check the possibility that fatigue or boredom might affect the results. In each class the 2 randomizations were randomly given to subjects. In all, 36 subjects took randomization 1, and 34 took randomization 2. The subjects took from 30 to 40 minutes to complete the experiment.

In order to make sure that this description is clear, let me give one example of a triad that the subject might see, in just the same form that he would see it. If it were triad number 7, it might be

7. Andy gets the book Someone gives Andy the book Andy loses the book.

Results. The statistic C (mean number of times classed together--see Experiment I) was calculated for all $\frac{11 \cdot 10}{2} = 55$ pairs of sentences for each of the 2 randomizations for each of the 2 classes. Pearson correlations were computed on the C values. The correlation between randomization 1 and randomization 2 for class 1 is r=.980 and for class 2 is r=.987. Thus the different randomizations didn't produce different results. The correlation between subjects from class 1 and subjects from class 2 for randomization 1 is r=.985 and for randomization 2 is r=.985. Thus the 2 classes produced similar results. From now on we will combine all the data and not consider the sub-groups. Table 1 lists the observed values of C for all the subjects.

Truax and I have tried a number of models on this data, making predictions for C from all of them. It is not unequivocally clear exactly which model is best, and their is no reason to go into all the analyses here. The models which work well all have much in common. They are hierarchies and the major (i.e., the first few) "cuts" are the same. A typical one of these models is presented as a tree in Figure 2. I chose this model to display because it also happens to be the hierarchy produced by a cluster analysis (Johnson, 1967, correctedness method) on the similarity matrix composed of the observed C values. The theoretical values of C predicted by this model are listed in parenthese in Table 3. The Pearson correlation between theory and data is r=.95.

The first thing to note about this model is its basic similarity to the hierarchical model on the words (Figure 1). The sentence contexts used changed the meanings so that we cannot expect, for example, the sentence "C gives A B" to appear in the same place in the sentence hierarchy as "give" appears in the word hierarchy. But the major features of meaning have retained their place. In the word analysis the first feature on which the verbs split was whether or not the (presumed) subject of the verb had the (presumed) object of the verb after time t. In the sentence analysis, the first split is on whether or not A has B after time t. The second feature in the word analysis is whether or not the subject has the object before time t. For the sentence analysis the second feature is whether or not A has B before time t.

As a side issue, I would like to speculate for a moment on what these results mean. The identification of A with the subject in the word experiment, and B with the object, suggests that in the word experiment subjects were providing their own frame, roughly of the form "A verb B". A curious fact is that Experiment I took subjects about 20 to 30 minutes to complete, while Experiment II took about 30 to 40 minutes, although there were almost twice as many triads in

Experiment II (165 to 84). Although we did not measure subjects' time, we do speculate that a sentence triad might be completed faster than a word triad, although there is much more reading to do for the sentence triad. It might be that the word triad takes longer because the subject has to provide for his own frame, which he is given in the sentence triad. Or it might simply be that the sentence triads are more interesting and thus are completed more quickly. In a pilot experiment for Experiment II we gave the same subjects 84 word triads and 84 sentence triads. It was our impression that they completed the sentence triads more quickly. Many of the subjects volunteered the information that the sentences were much easier and more fun than the words.

It may be that verbs can only be understood in a sentence context and thus subjects are forced to provide their own when not given one. This is compatable with the notion that a proper semantic analysis of a verb is as a predicate on arguments (e.g., Lakoff and Ross, 1967). A noun, on the other hand, is an elementary object in the logical system and doesn't have any arguments to be specified. Thus a subject should not be forced to create a sentence frame in a triad experiment on nouns and thus sentence triads should take longer than noun triads. This experiment has not yet been done.

Getting back to the main point of the experiment, it is clear that a hierarchical model for the semantic structure of the sentences does well. What about Bendix' paradigmatic model? We took Bendix' (p. 76) componential analysis seriously and considered it as a grid model, as in the kinship work I discusses earlier. There is no need to go into the details here since we are not primarily interested in discussing

Bendix' work (see Wexler and Truax, in preparation) for the details, and a discussion of how Bendix' method produced (to us) unsatisfactory results. For a linguistic criticism see Filmore (1969). Table 3 is Bendix' componential analysis, which we took as our grid model,

Insert Table 3 here

assuming, as in the kinship case, equal weights. We also had to add an analysis for "A has B", which does not appear in the table. This was done on the basis of Bendix' discussion of that form. When the process model is applied to this grid model predictions for the triads and for C values are obtained. The Pearson correlation between theory and data for the C values is r=0.58, which is rather lower than the r=0.95 for the hierarchical model. Inspection of the data shows many unsatisfactory predictions.

In an attempt to improve the predictions we modified Bendix' model in various ways, while trying to leave the basic analysis unchanged. The best correlation we could produce in this way was r=0.68. It is simply the case that some of Bendix' features, from a cognitive or psychological viewpoint at least, are not correct. The worst offender is the one relating to the before time t dimension, which Bendix marks positive if any relation exists between various objects before time t. Our experiment shows that, in fact, the feature is positive only if A has B before time t. Intuitively this also seems more unsatisfactory.

Although hierarchical models do well as semantic structures for both the words and sentences and Bendix' paradigmatic model does not do nearly so well, still it is not clear that another paradigmatic model would not predict the triad data as well as the hierarchical model. In fact, John Boyd and I have shown that for any hierarchy (more generally for any embedding structure) there exists a componential analysis (that is, a paradigmatic analysis on 0,1 values) which makes the same predictions on the triad test as does the hierarchy.

In general the componential analysis will have a very large number of dimensions, and the dimensions will not be intuitively explicable. Nevertheless the model does exist and we would like more concrete evidence distinguishing the models. This might be provided by latency (response time) measurements. We offer the following speculations as a natural model for the latencies of responses to triad tests on embedding structures. First, consider a strictly paradigmatic model. The process model says that the subject takes each pair of words, computes the difference of the values for each dimension in the model, sums these differences, and then chooses to class together that pair whose sum is smallest. The essential point here is that, no matter what the triad, the subject looks at all dimensions, and thus has the same amount of computing to do. Thus the non-hierarchical model predicts that each triad will take the same amount of time, that is, latencies will be equal for all triads.

Now consider the hierarchical structure model. What we said above for the non-hierarchical model will be true for the first depth of the embedding model. That is, no matter what the triad, there will be a constant time for processing of the structure at depth 1. At this point, however, a difference can arise. If there is some difference at the first depth, a decision is made there, and there is no further processing. If, on the other hand, all the first depth values are the same, the subject goes on to process the next depth. Processing continues until a difference is found. For now, we can make the very rough assumption that each depth takes the same amount of time to process. Thus, if r is the amount of time needed to process one level and L_{ijk} is the latency for a triad (ijk), we have

$$L_{ijk}=r\delta(ijk)+C,$$

where C is a constant and δ is the level at which a difference first appears for that triad.

The crucial point here is not the exact equation, since our assumption that r is a constant could be quite wrong. Rather, what is important is the prediction that different triads have different latencies, and that the greater the depth to which the subject has to process, the greater the response time.

These notions were tested in Experiment III, in which response times were measured for each triad.

Experiment III

Subjects. There were 60 subjects, who participated for required credit for an undergraduate psychology class. Subjects were run individually.

Procedure. The experiment was run on a Tachistoscope controlled by an Iconix control unit. The subject sat looking into the tachistoscope, which was dark. To start a trial the experimenter pressed a button. A buzzer sounded, the light came on, and the 3 words of the triad appeared to the subject. The words appeared for 9 seconds. In front of the subject were 3 buttons, labelled 1,2,3. The subject was instructed to select the word most different in meaning from the other 2 and then to push the appropriate button, #1 (the left-most) for the left-most word,

#2 (the middle one) for the middle one, and #3 (the right-most) for the right-most word. The subject's response was recorded on paper tape and also the time at which he made the response, measuring from when the words appeared. If the subject did not make a response in 9 seconds, no response was recorded. At the end of 9 seconds, the light went off. The experimenter changed cards to present the next triad and pressed a button to start the next trial. Subjects were told to respond as quickly as possible without error.

Triads. There are 6 ways in which a triad might be ordered from left to right. Six sets of cards for the tachistoscope were prepared, each containing all 84 triads. But the left to right order of the triads differed in each set. That is, a given left to right order for a given triad appeared in exactly one set. The sets were also randomized so that a given word appeared in first, second or third position approximately the same number of times in a set. A subject received one of the 6 set of cards. The cards in a set were shuffled after each subject so that the order of presentation of triads was randomized.

Results. A number of times subjects did not respond within the 9 second limit, and no response was recorded. In order to have data only for subjects who responded to most of the triads, the data from subjects who did not respond to at least 90% of the triads was discarded. This left us with 43 subjects. Only data from these subjects will be considered. Each triad had a total of 37 to 43 responses. The mean number of responses to a triad was 40.40.

The C values are given in Table 1. Comparison of these figures with those from Experiment I shows that the results are quite similar. The Pearson correlation between C values for Experiment I and Experiment III is r = 0.95. In general there seems to be a little more "noise" in this experiment. That is, the observed values of C are a little higher in this experiment for those pairs for which the model predicts that C=0. Showing this increase quite noticably is the pair of "opposites", "lend" and "borrow". It might be that the increased pressure to respond quickly in the experimental situation caused subjects to not use a semantic model, but rather to do something more akin to free association, which would put "lend" and "borrow" quite close. At any rate the triad choices in Experiment III are similar enough to those in Experiment I to allow us to proceed to the response times, which were the point of the experiment.

The mean reaction time over subjects for a triad varied from 3.726 seconds to 5.281 seconds. The mean overall triads and subjects was 4.731 seconds. In order to look at the predictions from the hierarchical model, we have to compute the level at which a first difference appears in the triad. We can do this from Figure 1. If 1 or 2 of the words is from the set [give,lend,lose] and the other word or two is from the other 6 words, then a difference appears at the first level. We call these "level 1" triads. Examples are [give,lose,have] and [get,take,lend]. There are 63 of these triads. If 1 or 2 words are from the set [have,keep] and the other word or two is from the set [get,take,find,borrow] then processing to the second level is necessary. These 16 triads are "level 2". We also consider the triad [give,lose,lend] to be level 2, on the assumption that the next level of structure (not shown) differentiates them. Clearly a binary tree structure could not do this, so there may be some higher level triads in this set. But in Wexler (1970) it is maintained that there is more than binary tree structure at this level, and that the level distinguishes all 4 verbs. At any rate, we assume that the 4 triads are level 3.

Let L_i, i=1,2,3, be the mean response time for triads of level i. Then we calculate $L_1=4.659$, $L_2=4.820$, $L_3=5.015$. A roughly linear trend may be discerned, as predicted from the hierarchical model. Figure 3 shows these values.

Insert Figure 3 here

An objection to the above calculation is that it includes the response times for all choices, even the ones not predicted by the hierarchical model. We do not know why these choices are made, but we certainly can't expect them to be made at the predicted "level". Thus, for levels 1 and 2, we performed a new calculation of the mean response times, including times only for the "correct", i.e., predicted choices. Call these values L_i' for level i. Then we calculate $L_1'=4.500$ and $L_2'=4.801$. For the difference between L_1' and L_2', $t=2.61$, $.01<p<.02$). In other words, the difference between level 1 and level 2 response times is even greater when only "correct" responses are considered. This calculation cannot be done for level 3 because we do not predict a choice here, and indeed, there is not one consistently selected word in each of these triads.

It is well-known that when choices in a reaction time experiment are more equi-probable the response times tend to be slower. The data show that the 3 words in level 3 triads are more equi-probale than in the level 1 or 2 triads. Thus this could explain, without recourse to semantic models, why level 3 triads are slower than the others. If it was also the case that level 2 triad choices were more equi-probable than the level 1 choices, than an artifact might explain all the results. We took as a measure of "dominance" of choice (the inverse of equi-probability) the proportion of responses which were the "correct" or predicted responses. Since in all but 4 of the 80 level 1 and 2 triads the predicted response was also the most frequent response, this measure is quite similar to the proportion of responses which are most frequent for each triad. This dominance measure turned out to be .706 for level 1 and .700 for level 2. These measures clearly are not different, especially if one compares them to a value of .519, which is the proportion of "most frequent" choices for lev 1 3 triads. Thus an artifact (at least this artifact) does not explain the results.

The results of Experiment III thus seem to show that triads which have to be processed more deeply according to the hierarchical model have longer response times.

Discussion

The 3 experiments provide some evidence that the structure of semantic domains is hierarchical. Experiment I showed that a hierarchical model fit the triad experiment for the set of "have" words. Experiment II extended these results to sentences instead of words. Experiment III suggested that response times behave the way a hierarchical model predicts they should.

I have alternated between describing the hierarchical structure in 2 different ways. One way is to simply write the tree, as in Figure 1.

The second way is to write a sequence of features for each word, as, for example, get=(0),(1). Now it is certainly true that these 2 representations (called the "tree" and "lexicon") provide exactly the same information about semantic features of a word. Thus from a purely linguistic (or "competence") point of view the 2 representations are identical. In fact, Katz and Fodor (1963) would list the redundancies implied by the tree in their dictionary before the lexicon. That is, they assume that the information represented is equivalent but that the tree is less redundant.

But for a moment I want to take the tree and lexicon representations seriously as psychological models. What demands are put upon the semantic model? It seems that given only the phonological form of the word, a speaker of the language has to be able to go directly to the semantic representation of the word. Of course there are important complications such as the role of context in disambiguating words with more than one meaning, but I will ignore these problems here). So I assume that the point of entry into the semantic representation of a word is exactly where the word appears in the model. Thus in the lexicon model (sequences of features) the point of entry is closest to the "higher-up features". In the tree model, on the other hand, the point of entry is closest to the "lower-down" features, and if the tree is deep, there may be a large number of features between the point of entry and the highest feature.

Consider the triad task. If semantic structure follows the lexicon model, then the hierarchical process model can be looked at as a direct information processing model, whose steps can actually be carried out as indicated, starting from the word and going on to the highest features. The process model looks at the highest features first. For the lexicon model these are right next to the words, which are the point of entry.

The next highest features are next in sequence, and the model proceeds directly. But for the tree model the case is not so simple. The lowest features are next to the point of entry. Thus to follow the process model, the subject would have to work up the tree to the first feature and mark it with the word of the triad that it dominated. For example, if the subject was working on "get" he would have to mark the left-most node immediately dominated by the root with "get". If the 3 words are all the same on the first level the subject would have to repeat the process again, up to the next level (unless he had marked all the nodes on his way up--but think of the memory burden for a deep tree). This rather tortuous model is quite inefficient, involving either much redundancy of processing or a large memory burden or both.

Of course the hierarchical process model we have given is not the only one that makes the correct predictions of choices for the tree model. Consider the following natural alternative. Let the "least upper bound" of 2 words be the lowest node in the tree which dominates the 2 words (where the root is the highest). Then there are 3 least upper bounds for a triad, one for each pair. It is easy to show that for a binary tree 1 of these least upper bound is dominated by the other 2. It is clear that this model makes exactly the same choice predictions as our first hierarchical process model. But note that the lower the "level" of the triad (that is the earlier, from the top down, a differnce appears), then the higher the least upper bounds, and thus the more processing that has to be done to make a decision. Thus the lower the level the greater the response time. But this is exactly opposite the prediction made by the first hierarchical process model, and, in fact, exactly opposite the response time results obtained in

Experiment III. Therfore this "bottom-up" model doesn't seem to capture cognitive processes correctly.

I haven't formally spelled out the details of these alternative models, and it is quite possible that variants of the models exist which would make correct predictions. Clearly much work, both theoretical and experimental, remains to be done before we understand semantic structure and process. The major point of this last discussion was to point out that lexicon and tree structures are not psychologically equivalent, and that the difference is amenable to experimental attack.

A lexicon representation is more redundant than a tree model because the same features have to be repeatedly listed for each word to which they apply whereas a tree model only lists each feature once. Thus it would seem (though this hasn't been spelled out rigourously) that the lexicon demands a much larger memory storage than would a tree representation. But we noted that the first model that we thought of to work on a tree was quite redundant in its operation. Thus there might be an interesting trade-off between memory storage and amount of processing. Although the lexicon demands greater amounts of memory, access to it might be more efficient than to a tree, especially if access to the "higher" features was demanded first. But this might be exactly the requirement of natural language processing. If I say "I borrowed the book", perhaps the first or most important feature of that sentence is that now "I have the book". The fact that I have an obligation to return the book (or something similar to this) is a lower-down feature and perhaps is important only if the context demands it.

But speculation runs away with me. Let me finish by noting some possible objections to the over-all thrust of this work, with my responses to them. Some of the questions may be naive, but I have been asked them.

1) If this is work in semantic memory then where are the memory experiments?

I have dealt with this question earlier in the paper. Clearly long-term memory is involved in the processes that I have described.

2) Aren't there individual differences in the responses to the triad experiment? How then can one structural model be offered for semantics?

Of course there are differences in responses, but this does not necessarily imply that the structure differs from subject to subject. The process that a subject uses in making his choice might differ. In fact, who knows why there are differences? For the "have" words I have not yet made any attempt to see if more than one pattern of response exists. This is a very difficult question. For a start on one approach for kinship terms see Wexler and Romney (1970).

3) How can you depend on a time difference of 200 or 300 milliseconds to distinguish various structures? Simply, cognitive processes are quick. That time is the time it takes to traverse a branch in a "feature" tree. Collins and Quillian (1969) measured the time it took to traverse a branch in a "superset" tree at 75 milliseconds.

4) Meaning is defined on sentences; words are not enough. This would be an appropriate comment if I claimed to have a general theory of semantic understanding. But I don't. None of this work bears on how the meaning of a sentence depends on the meaning of the words in the sentence and the syntax of the sentence. But Experiment II shows that the experimental technique might be useful for sentences also.

5) How about other ways of measuring similarity of words? I have done a sorting experiment on 62 "have" words, including the 9 studies here. Examples of the other words are "receive", "donate", etc. The major features of the semantic analysis given here appear to replicate nicely for that experiment.

6) The triad test is far too simple to represent a speaker's knowledge of the meanings of words. But it is important to distinguish between an experimental tool and a theory of cognition. Consider linguistics. The basic datum for the study of syntax is simply a decision about whether a sentence is grammatical or not, that is a 0-1 decision. (There are a few other data used, such as whether 2 sentences have the same meaning, but these are not nearly so common as the grammaticality judgment, and they are themselves simple). Nevertheless, elaborate theories of grammar are formulated to explain this simple datum. In fact the triad test is formally more complicated than the judgment of grammaticality, since 3 items are involved rather than 2. Thus there is no reason that interesting theory can't be built around simple, even "trivial" judgments. Let me close with the following quote from Weinreich (1962, p. 25):

> "As speakers of English, we can state with little hesitation that in each of the following triplets, two words belong more closely together than a third. . . . To give an explicit account of such intuitions is a good way of beginning descriptive semantics."

References

Bendix, E. H. Componential analysis of general vocabulary: the semantic structure of a set of verbs in English, Hindi and Japanese. Bloomington: Indiana University, 1966. Also *International Journal of American Linguistics*, 32, part 2.

Bower, G.H. Organization and Memory. *Cognitive Psychology*, 1970, *1*.

Boyd, J.P. and Chalmers, D.K. Paper presented at the Psychonomic Society Meetings, 1967.

Collins, A., and Quillian, R., Retrieval time fron semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1969, *8*.

Fillmore, C. Review of Bendix, 1966. Department of LInguistics, Ohio State University, Technical Report, 1969.

Katz, J.J., and Foder, J.A. The structure of a semantic theory. *Language*, 1963, *39*, 170-210.

Kintsch, W., Miller, R.F., and Hogan, R.M. Semantic structure and clustering in free recall. Delivered at Workshop on Cognitive Organization and Psychological Brocesses, Huntington Beach, California, August, 1970.

Johnson, S.C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241-253.

Lakoff, G. and Ross, J.R., Is deep structure necessary? Duplicated, M.I.T., 1967.

Miller, G.A. A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 1969, *6*, 169-191.

Romney, A.K. and D'Andrade, R.G. Cognitive aspects of English kin terms. In A.K. Romney and R.G. D'Andrade (Eds.), *Transcultural studies in cognition*, (*American Anthropologist*, 66:3, part 2). Menasha, Wisconsin, 1964, 146-170.

Wexler, K.N. Embedding structures for semantics. University of California, Irvine, 1970 (mimeographed).

Wexler, K.N. and Romney, A.K. Individual variations in cognitive structures. Presented at the Conference on multidimensional scaling in the behavioral sciences, June, 1969. To appear in Shepard, R.N., Romney, A.K. and Nerlove, S.B. (Eds.), *Recent advances in multi-dimensional scaling*.

Weinreich, U., Lexicographic definition in descriptive semantics, in *Problems in Lexicography*, Fred W. Householder and Sol Saporta, eds. (Bloomington, Indiana University Research Center in Anthropology, Folklore, and Linguistics, Publication 21, 1962), 25-43.

Footnotes

1. Dr. Arnold Binder kindly allowed his laboratory to be used for Experiment III. I am also indebted to Dorothy Martin for running and helping to analyze Experiments I and III, and to Werner Karle for help with the equipment for Experiment III.
2. For the case of binary trees, the process model in Wexler (1970) is identical with that of Geohegan. Embedding structures, under the name "compnential trees" and via a recursive partition definition were first introduced independently of me by Boyd and Chalmers (1967). Miller (1969) has studied hierarchies by use of the sorting method.

Table 1. Theoretical and Observed Values of E(C) = mean number of times a pair of verbs is classed together

	Observed		Theoretical Predictions	
	Exp. I	Exp. III	Paradigmatic Model	Hierarchical Model
lend-give	6.12	6.16	6.33	6.33
have-keep	6.05	6.06	7.00	7.00
get-take	5.42	4.79	5.67	5.67
get-borrow	5.08	4.40	5.67	5.67
get-find	4.91	4.75	5.67	5.67
borrow-take	4.80	4.36	5.67	5.67
give-lose	4.45	3.44	6.33	6.33
find-borrow	4.35	2.56	5.67	5.67
lend-lose	4.25	4.24	6.33	6.33
find-take	4.22	3.89	5.67	5.67
keep-take	4.12	3.14	1.50	3.00
find-keep	3.42	3.17	1.50	3.00
have-take	2.96	2.25	1.50	3.00
find-have	2.69	2.93	1.50	3.00
get-keep	2.66	2.67	1.50	3.00
get-have	2.46	2.99	1.50	3.00
borrow-keep	2.21	1.69	1.50	3.00
borrow-have	2.08	2.09	1.50	3.00
lend-borrow	1.21	2.63	0	0
get-give	1.12	1.25	0	0
lend-have	.92	1.26	2.00	0
give-keep	.92	.67	2.00	0
lend-get	.83	1.36	0	0
find-lose	.83	1.39	0	0
borrow-give	.75	1.76	0	0
find-give	.67	.69	0	0
lend-take	.67	1.88	0	0
lend-find	.58	.55	0	0
give-have	.58	1.20	2.00	0
borrow-lose	.50	.99	0	0
lose-keep	.46	.52	2.00	0
get-lose	.46	.43	0	0
lend-keep	.46	.58	2.00	0
give-take	.42	1.46	0	0
lose-take	.33	.63	0	0
lose-have	.28	.57	2.00	0

Table 2. A Crude Semantic Model

		after time t	
		Have	Not Have
before time t	Have	have keep	give lend lose
	Not Have	get take borrow find	

Table 3. Theoretical and Observed Values of E(C)

for Experiment II

		Observed	Theoretical
1.	loses/rid of	8.40	9.00
2.	C lends/borrows	8.03	9.00
3.	keeps/has	7.39	9.00
4.	C gets/C gives	7.37	9.00
5.	C gives/C lends	6.26	7.00
6.	C gets/C lends	6.16	7.00
7.	takes/borrows	5.99	6.00
8.	C gives/borrows	5.53	7.00
9.	finds/gets	5.37	9.00
10.	C gives/takes	5.27	6.00
11.	C gets/borrows	5.19	7.00
12.	gets/C gets	5.14	4.00
13.	gets/takes	4.97	4.00
14.	C gets/takes	4.87	6.00
15.	gets/C gives	4.74	4.00
16.	gets/borrows	4.57	4.00
17.	C lends/takes	4.51	6.00
18.	keeps/finds	3.97	2.00
19.	gets/C lends	3.89	4.00
20.	has/finds	3.82	2.00
21.	keeps/gets	3.69	2.00
22.	has/gets	3.37	2.00
23.	finds/borrows	3.02	4.00
24.	finds/takes	3.00	4.00
25.	finds/C gives	2.72	4.00
26.	keeps/C gives	2.64	2.00
27.	has/C gets	2.64	2.00
28.	finds/C lends	2.60	4.00
29.	finds/C gets	2.54	4.00
30.	has/C gives	2.50	2.00
31.	has/C lends	2.43	2.00
32.	keeps/takes	2.42	2.00
33.	has/takes	2.42	2.00
34.	keeps/C gets	2.23	2.00
35.	keeps/borrows	2.19	2.00
36.	has/borrows	2.16	2.00
37.	keeps/C lends	1.93	2.00
38.	loses/finds	1.64	0
39.	rid of/keeps	0.77	0
40.	loses/keeps	0.74	0
41.	loses/has	0.74	0
42.	rid of/finds	0.61	0
43.	loses/gets	0.59	0
44.	rid of/gets	0.53	0
45.	rid of/takes	0.49	0
46.	rid of/has	0.46	0
47.	rid of/borrows	0.37	0
48.	loses/borrows	0.36	0
49.	loses/C lends	0.33	0
50.	loses/takes	0.33	0
51.	rid of/C gives	0.26	0
52.	loses/C gets	0.23	0
53.	rid of/C gets	0.23	0
54.	loses/C gives	0.21	0
55.	rid of/C lends	0.20	0

Note: The sentences are abbreviated. Thus pair #1 is "A loses B," "A gets rid of B". Pair # 11 is "C gets A B, A borrows B from C."

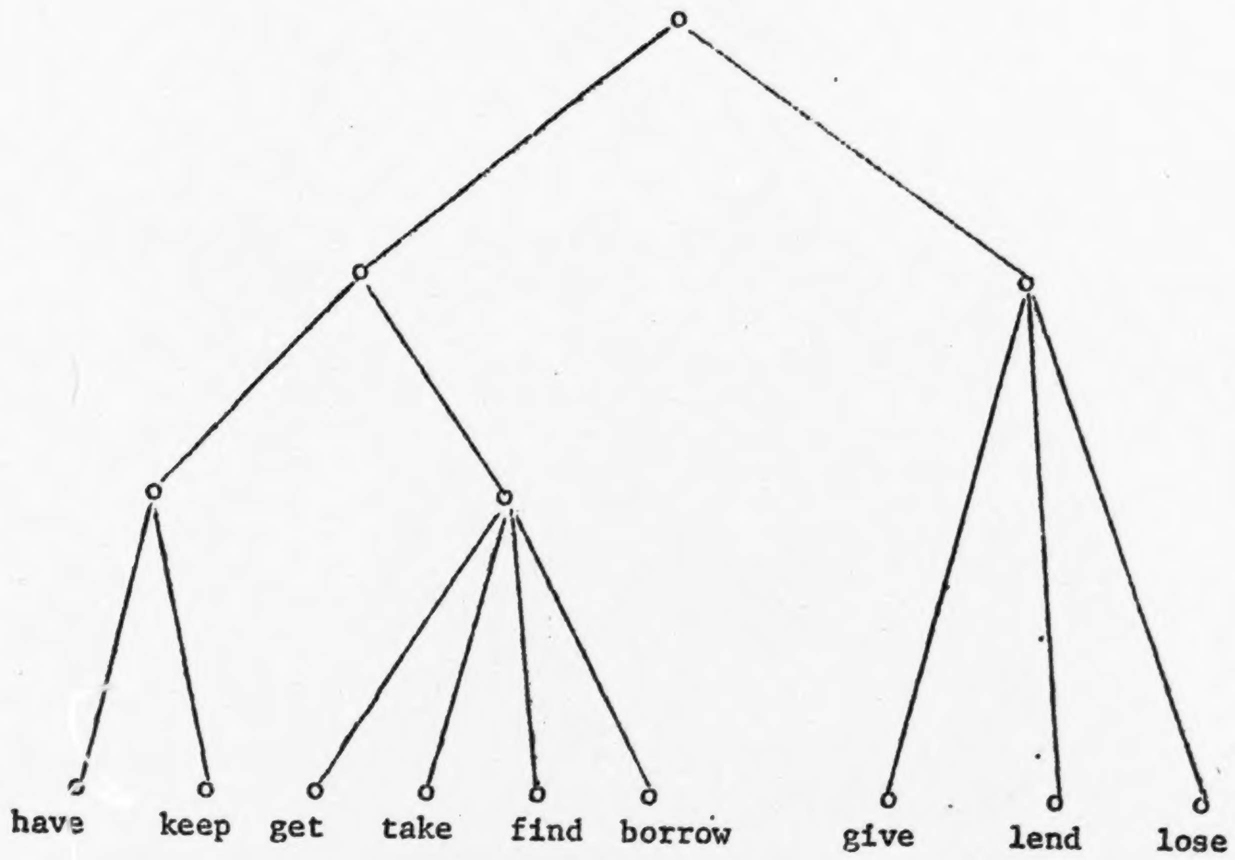


Figure 1

A Hierarchical Model

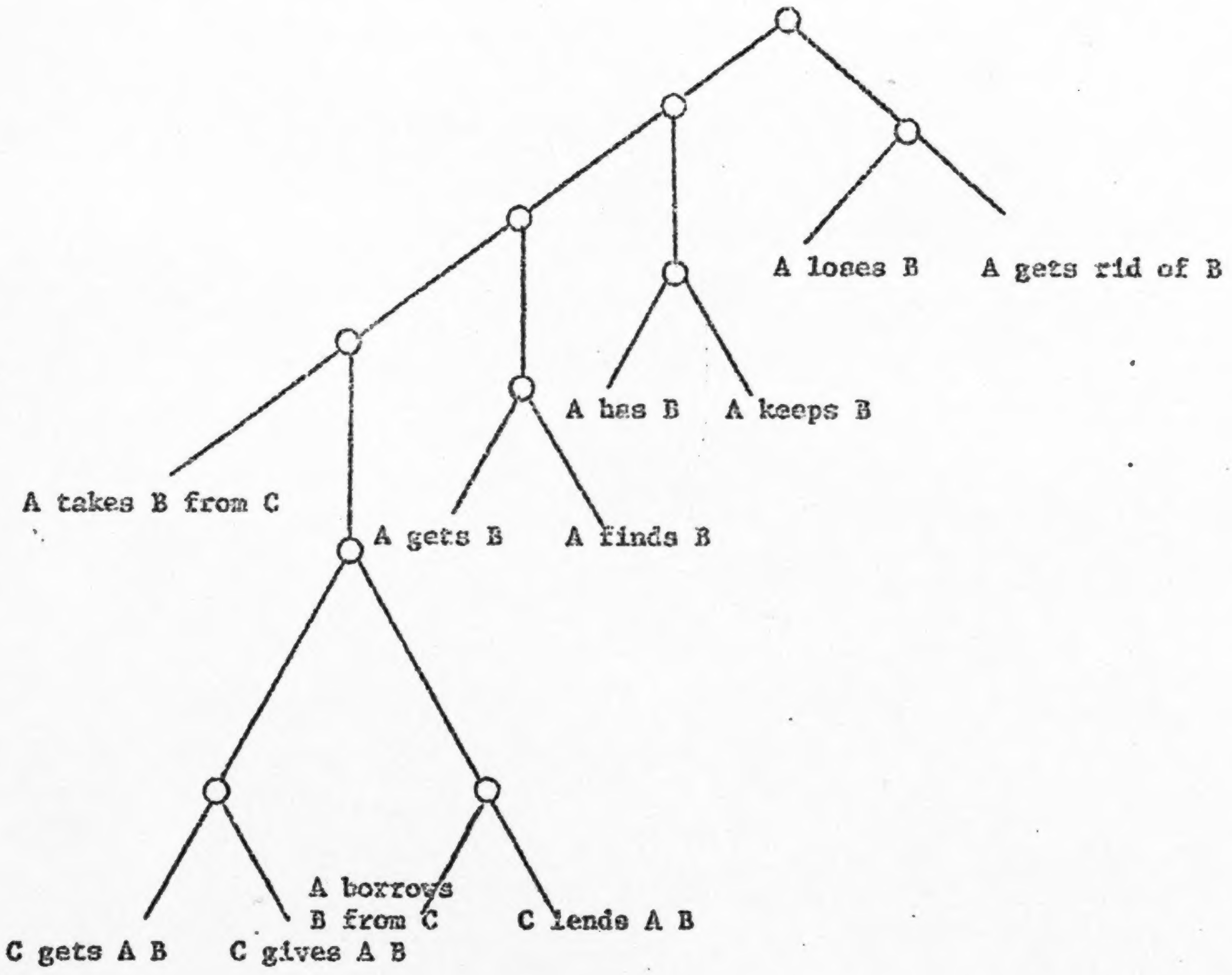


Figure 2. A Model for the Meaning of Sentences

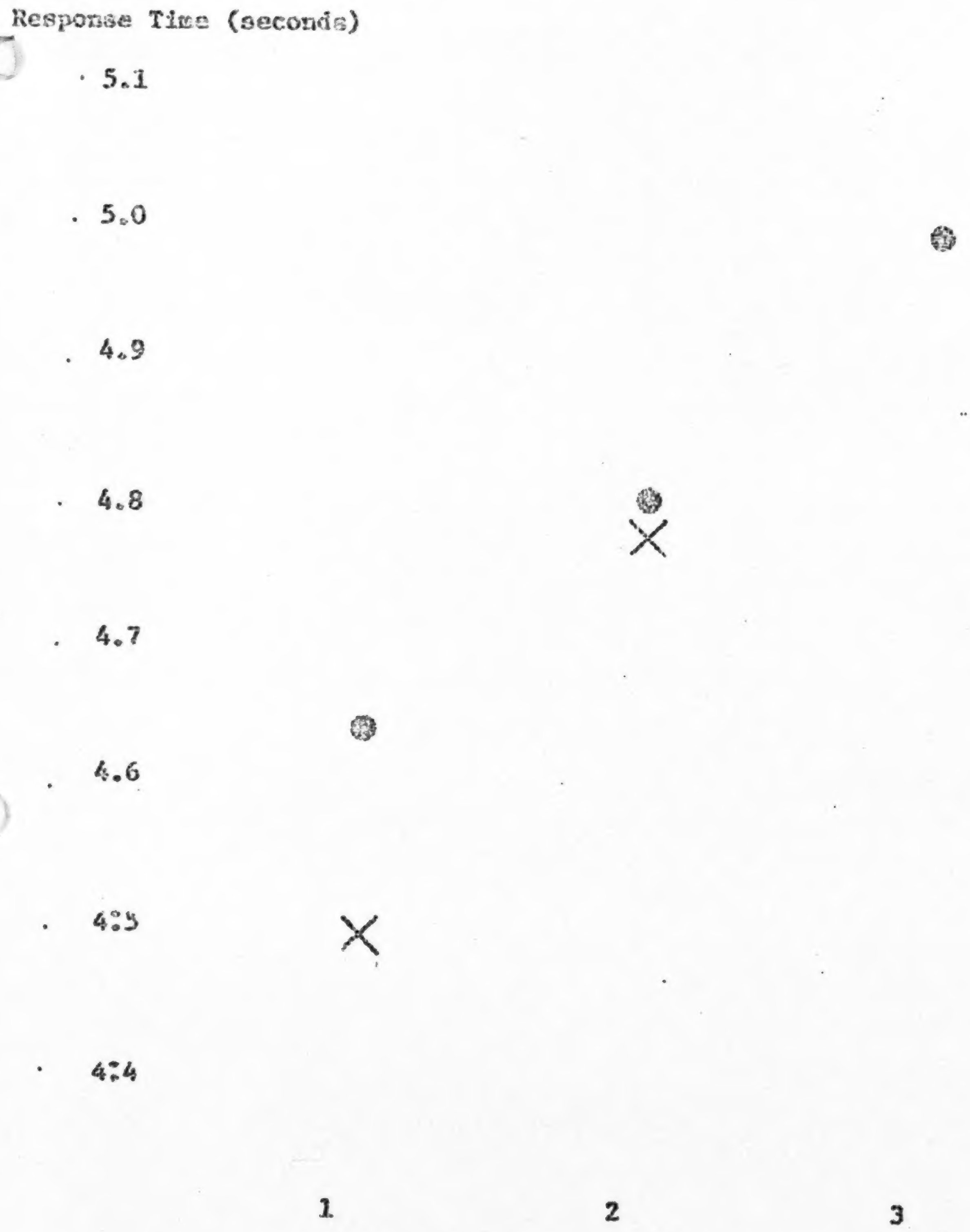


Figure 3. Mean Response Time per Level in Experiment III.

● L (all responses)
× L' (predicted responses only)